Written Testimony
of
Sam Altman
Chief Executive Officer
OpenAI
Before the U.S. Senate Committee on the Judiciary
Subcommittee on Privacy, Technology, & the Law

Chairman Blumenthal, Senator Hawley, and members of the Judiciary Committee, thank you for the opportunity to testify today about large neural networks. I am Sam Altman, Chief Executive Officer of OpenAI, a company that studies, builds, and deploys artificial intelligence (AI) and has created AI tools such as ChatGPT, Whisper, and DALL·E 2. OpenAI was founded on the belief that safe and beneficial AI offers tremendous possibilities for humanity. I am grateful for the opportunity to speak about our experiences developing cutting-edge AI technology and studying AI safety, and our interest in working collaboratively with governments to ensure the development and widespread availability of safe and beneficial AI tools. We believe it is essential to develop regulations that incentivize AI safety while ensuring that people are able to access the technology's many benefits.

About OpenAI

OpenAI is a San Francisco based company created in 2015 to ensure that artificial general intelligence (AGI) benefits all of humanity. OpenAI's work is driven by our charter, in which we commit to working toward the broad distribution of the benefits of AGI, to maximizing the long term safety of AI systems, to cooperating with other research and policy institutions, and to serving as a technical leader in AI to accomplish these objectives.

OpenAI has an unusual structure that ensures that it remains focused on this long-term mission. We have a few key economic and governance provisions:

- First, the principal entity in our structure is our Nonprofit, which is a 501(c)(3) public charity.
- Second, our for-profit operations are subject to profit caps and under a subsidiary that is fully controlled by the Nonprofit.
- Third, because the board serves the Nonprofit, each director must perform their fiduciary duties in furtherance of its mission—safe AGI that is broadly beneficial. While the for-profit subsidiary is permitted to make and distribute profit, it is subject to this mission. The Nonprofit's principal beneficiary is humanity, not OpenAI investors.
- Fourth, the board remains majority independent. Independent directors do not hold equity in OpenAI.
- Fifth, profit for investors and employees is capped by binding legal commitments. The Nonprofit retains all residual value for the benefit of humanity.

This structure enables us to prioritize safe and beneficial AI development while helping us secure the necessary capital to develop increasingly powerful AI models. For example, in January, Microsoft announced a multiyear, multibillion dollar investment in OpenAI, following previous investments in 2019 and 2021.[1] This investment provides necessary capital and advanced supercomputing infrastructure for OpenAI to develop, test, and improve our technology. Microsoft is an important investor in OpenAI, and we value their unique alignment with our values and long-term vision, including their shared commitment to building AI systems and products that are trustworthy and safe. At the same time, OpenAI remains an entirely independent company governed by the OpenAI Nonprofit. Microsoft has no board seat and does not control OpenAI. Furthermore, AGI technologies are explicitly reserved for the Nonprofit to govern.

OpenAI Technology and Tools

OpenAI is a leading developer of large language models (LLMs) and other AI tools. Fundamentally, the current generation of AI models are large-scale statistical prediction machines – when a model is given a person's request, it tries to predict a likely response. These models operate similarly to auto-complete functions on modern smartphones, email, or word processing software, but on a much larger and more complex scale.[2] The model learns from reading or seeing data about the world, which improves its predictive abilities until it can perform tasks such as summarizing text, writing poetry, and crafting computer code. Using variants of this technology, AI tools are also capable of learning statistical relationships between images and text descriptions and then generating new images based on natural language inputs.

Our models are trained on a broad range of data that includes publicly available content, licensed content, and content generated by human reviewers.[3] Creating these models requires not just advanced algorithmic design and significant amounts of training data, but also substantial computing infrastructure to train models and then operate them for millions of users.

Our major recent releases include tools that can generate images and text. In early 2022, we launched a research preview of DALL·E 2, an AI system that can create realistic images and art from a description in natural language.[4] Millions of users are now creating and improving images using DALL·E and sharing their creations with the world. Since the initial preview, we have expanded DALL·E's capabilities, including launching a DALL·E Application Programming Interface (API) to help developers integrate DALL·E into apps and products.[5]

---

[1] "Microsoft and OpenAI Extend Partnership." *Microsoft*, 23 Jan. 2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.
[2] Johnson, Steven. "A.I. Is Mastering Language. Should We Trust What It Says?" *New York Times Magazine*, 15 Apr. 2022, https://www.nytimes.com/2022/04/15/magazine/ai-language.html.
[3] "Our Approach to AI Safety." *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety.
[4] "DALL·E 2." *OpenAI*, https://openai.com/product/dall-e-2. Accessed 14 May 2023.
[5] "DALL·E API Now Available in Public Beta." *OpenAI*, 3 Nov. 2022, https://openai.com/blog/dall-e-api-now-available-in-public-beta.

On the text side, we have trained and publicly released a number of LLMs, beginning with the GPT-2 family of models in 2019[6] and the GPT-3 family of models in 2020.[7] In November 2022, we released ChatGPT.[8] These models can be used to organize, summarize, or generate new text. They "understand" user queries and instructions, then generate plausible responses based on those queries. The models generate responses by predicting the next likely word in response to the user's request, and then continuing to predict each subsequent word after that. The models are available for free in most of the world; we also have launched a pilot subscription service, ChatGPT Plus, that provides additional benefits to users,[9] and we make the models available as an API for developers to build applications and services.

In March of this year, we released GPT-4, our most advanced system, which is capable of producing more useful, more creative, more collaborative, and more accurate outputs than previous OpenAI products.[10] GPT-4 is available on ChatGPT Plus and (as with other GPT models) as an API for developers to build applications and services.

AI Continues to Improve People's Lives

OpenAI's mission is to ensure that AI systems are built, deployed, and used safely and beneficially. We see firsthand both the potential and the actual positive impact that these systems have in improving people's lives. We hear from users all over the world about how much they love our tools and how AI helps them in many ways, including helping them write complex computer code more quickly, enhancing their productivity and creativity, augmenting their existing strengths, helping them learn new skills, and expanding their businesses.[11] Here are just some of the ways that customers are using our products:

- Educational non-profit Khan Academy is piloting a program that uses GPT-4 to power a personalized virtual tutor for students and a classroom assistant for teachers.[12]
- Morgan Stanley is using GPT-4 to power an internal-facing chatbot that performs a comprehensive search of wealth management content and "effectively unlocks the cumulative knowledge of Morgan Stanley Wealth Management," helping their financial advisors better serve their clients.[13]

---

[6] "GPT-2: 1.5B Release." *OpenAI*, 5 Nov. 2019, https://openai.com/research/gpt-2-1-5b-release.
[7] "OpenAI API." *OpenAI*, 11 June 2020, https://openai.com/blog/openai-api.
[8] "Introducing ChatGPT." *OpenAI*, 30 Nov. 2022, https://openai.com/blog/chatgpt.
[9] "Introducing ChatGPT Plus." *OpenAI*, 1 Feb. 2023, https://openai.com/blog/chatgpt-plus.
[10] "GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses." *OpenAI*, https://openai.com/product/gpt-4. Accessed 14 May 2023. Vincent, James. "OpenAI Announces GPT-4 — The Next Generation of Its AI Language Model." *The Verge*, 14 Mar. 2023, https://www.theverge.com/2023/3/14/23638033/openai-gpt-4-chatgpt-multimodal-deep-learning.
[11] Paris, Francesca and Larry Buchanan. "35 Ways Real People Are Using A.I. Right Now." *New York Times: The Upshot*, 14 Apr. 2023, https://www.nytimes.com/interactive/2023/04/14/upshot/up-ai-uses.html.
[12] "Khan Academy." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/khan-academy.
[13] "Morgan Stanley." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/morgan-stanley.

- Stripe is using GPT-4 in a variety of ways, including improving its custom support operations, helping to answer support questions about technical documentation, and helping to detect fraud on its community platforms.[14]
- Harvey, a platform for legal professionals, is using GPT-4 to make tasks such as research and drafting more efficient so they can focus more time on strategy, and deliver a higher quality service to more clients.[15]
- Speak, the fastest-growing English application in South Korea, is using Whisper, our automatic speech recognition AI system, to power an AI speaking companion and provide true open-ended conversational practice.[16]
- Weave is using our tools to build a collaboration platform for scientists, specifically focused on breakthroughs in oncology.
- Creative professionals from movie directors to indie musicians are using our image generation tool, DALL·E, to augment their creative processes, from rapid storyboarding to creating cover art that would not have previously been possible.

We also partner with nonprofit and other organizations to explore socially beneficial uses of our tools. For example, our technology enables a nonprofit called Be My Eyes to help people who are blind or have low vision use our models to help describe what they are seeing. These users normally rely on volunteers for help with hundreds of daily life tasks, and we're seeing that an AI powered system called "Virtual Volunteer" can help reach the same level of context and understanding as a volunteer.[17] The Trevor Project has used GPT-2 to significantly scale its efforts to prevent suicide among LGBTQ teens, while Lad in a Battle has used DALL·E to bring joy to pediatric cancer patients. The government of Iceland is using GPT-4 in its preservation efforts for the Icelandic language,[18] and other countries have expressed interest in using this same model to preserve under-resourced languages.

I also want to share the story of Ben Whittle, a pool installer and landscaper with dyslexia. Ben feared that his dyslexia would harm his email communications with his clients. One of Ben's clients created an AI tool built on top of our technology to help Ben write better emails by making suggestions, improving his grammar, and adding professional niceties. Ben now uses this AI tool for all his work emails and believes it played a significant role in securing a $260,000

---

[14] "Stripe." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/stripe.
[15] "Introducing Our First Investments." *OpenAI Startup Fund*, 1 Dec. 2022, https://openai.fund/news/introducing-our-first-investments. "A&O Announces Exclusive Launch Partnership with Harvey." *Allen & Overy*, 15 Feb. 2023, https://www.allenovery.com/en-gb/global/news-and-insights/news/ao-announces-exclusive-launch-partnership-with-harvey.
[16] "Introducing Our First Investments." *OpenAI Startup Fund*, 1 Dec. 2022, https://openai.fund/news/introducing-our-first-investments.
[17] "Be My Eyes." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/be-my-eyes.
[18] "Government of Iceland." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/government-of-iceland.

contract for his company. Ben was quoted in the Washington Post, saying, "this has given me exactly what I need."[19]

This is just one way our technology can benefit people as they learn to adopt and use AI tools. These opportunities are why former U.S. Treasury Secretary Lawrence Summers has said that AI tools such as ChatGPT might be as impactful as the printing press, electricity, or even the wheel or fire.[20]

We feel an immense amount of excitement, opportunity, and responsibility in being involved with helping build the future.

AI Safety Practices

While we believe the benefits of the tools we have deployed vastly outweigh the risks, ensuring their safety is vital to our work and we make significant efforts to ensure safety is built into our systems at all levels. In the sections below, I discuss our general approach to safety and some of the specific steps we take to make our models safer.

Prior to releasing each new version of our models, OpenAI conducts extensive testing, engages external experts for feedback, improves the model's behavior with techniques like reinforcement learning from human feedback (RLHF), and implements safety and monitoring systems.[21]

The release of our latest model, GPT-4, provides an illustrative example. After we developed GPT-4, we spent more than 6 months evaluating, testing, and improving the system before making it publicly available.[22] In addition to our own evaluations, we engaged with external AI safety experts in a process known as “red teaming,” through which they helped identify potential concerns with GPT-4 in areas including the generation of inaccurate information (known as “hallucinations”), hateful content, disinformation, and information related to the proliferation of conventional and unconventional weapons.[23] This process helped us to better understand potential usage risks and ways to address those risks.

In each of these areas, we developed mitigations to increase safety in significant ways. Some of our work involved making adjustments to the data used to train the model, during what is called

---

[19] Harwell, Drew, Nitasha Tiku, and Will Oremus. “Stumbling with Their Words, Some People Let AI Do the Talking.” *Washington Post*, 10 Dec. 2022, https://www.washingtonpost.com/technology/2022/12/10/chatgpt-ai-helps-written-communication/.
[20] Summers, Lawrence. *Bloomberg: Wall Street Week*, 9 Dec. 2022, https://www.youtube.com/watch?v=iR31ZAacyGM.
[21] “Our Approach to AI Safety.” *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety.
[22] “Our Approach to AI Safety.” *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety.
“GPT-4 System Card.” *OpenAI*, 23 Mar. 2023, p. 19, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[23] GPT-4 System Card.” *OpenAI*, 23 Mar. 2023, p. 4, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

the pre-training stage. Other interventions took place after initial training of the model.[24] At the pre-training stage, for example, we reduced the quantity of erotic text content in our dataset.[25] After the pre-training stage, our primary method for shaping GPT-4's behavior involves having people provide feedback on model responses, in order to help teach our models to respond in a way that is safer and more useful.[26] We also teach the model to try to refuse harmful requests and to respond more appropriately in the face of sensitive requests. These efforts empirically reduced the likelihood that the model would generate harmful or inaccurate content.[27] When asked to generate disallowed content (as defined by our usage policies), GPT-4 refuses to do so more than 99% of the time.[28] While our models still have limitations and can generate disallowed or inaccurate information in some cases, we've made significant progress through these safety efforts, and we're continuing to build on them.

Deployment Safety and Learning

We work hard to understand and prevent risks before deployment.[29] However, we can't anticipate every beneficial use, potential abuse, or failure of the technology. This is in large part because these systems are still human-directed—they try to follow user instructions to carry out tasks. Learning from and responding to real-world use by actual people is vital for creating safer AI systems..[30]

Our deployment practices involve cautiously and gradually releasing new AI models—with substantial safeguards in place—to gradually larger groups of people, making continuous improvements based on the lessons learned. We also make our most capable models available through our own services (and through an API), which allows us to monitor for and take action on misuse, and continually build mitigations that respond to the real ways people misuse our systems.

As described in our Usage Policies, OpenAI expressly prohibits the use of its tools for certain activities, including but not limited to, generation of violent content, malware, fraudulent activity, high-volume political campaigning, and many other unwelcome areas.[31]

---

[24] "GPT-4 System Card." *OpenAI*, 23 Mar. 2023, pp. 21-22, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[25] GPT-4 System Card." *OpenAI*, 23 Mar. 2023, pp. 3, 21, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[26] Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *arXiv*, 4 Mar. 2022, https://arxiv.org/pdf/2203.02155.pdf.
[27] GPT-4 System Card." *OpenAI*, 23 Mar. 2023, p. 20, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[28] GPT-4 System Card." *OpenAI*, 23 Mar. 2023, pp. 22-24, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[29] "Lessons Learned on Language Model Safety and Misuse." *OpenAI*, 3 Mar. 2022, https://openai.com/research/language-model-safety-and-misuse.
[30] "Our Approach to AI Safety." *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety.
[31] "Usage Policies." *OpenAI*, 23 Mar. 2023, https://openai.com/policies/usage-policies.

We use a combination of automated detection systems and human review to detect potentially violating behavior in order to warn users or take enforcement actions. We use our newest models to help to identify unsafe content—this reduces the need for human moderators to be exposed to harmful or explicit content, helps us to quickly refine our moderation policies, and reduces the time needed to build safety tools. We also provide a free suite of moderation and safety tools to our developers to integrate into their products.

We strive to be agile and responsive to customer concerns. We are continuously updating and improving our models and products based on feedback from our customers, users, the public, and other stakeholder groups, including governments.

Iterative deployment has other advantages for AI safety. We believe that people and our institutions need time to update and adjust to increasingly capable AI, and that everyone who is affected by this technology should have a significant say in how AI develops further. The public dialogue on generative AI has advanced dramatically since OpenAI launched ChatGPT last November. Iterative deployment has helped us bring various stakeholders into the conversation about the adoption of AI technology more effectively than if they hadn't had firsthand experience with these tools.[32]

Privacy

OpenAI takes the privacy of its users seriously and has taken a number of steps to facilitate transparent and responsible use of data. First, we don't use any user data to build profiles of people for the purposes of advertising, promoting our services, or selling data to third parties. We also do not use data submitted by customers via our API to train or improve our models, unless customers explicitly ask us to do this. We may use ChatGPT conversations to help improve our models, but we provide users with several ways to control how their conversations are used. Any ChatGPT user can opt-out of having their conversations be used to improve our models.[33] Users can delete their accounts,[34] delete specific conversations from the history sidebar, and disable their chat history at any time.[35]

While some of the information we use to train our models may include personal information that is available on the public internet, we work to remove personal information from the training dataset where feasible, teach our models to reject requests for personal information of private individuals, and respond to requests from individuals to remove their personal information from our systems. These steps reduce the likelihood that our models might generate responses that include the personal information of private individuals.

---

[32] "Our Approach to AI Safety." *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety.
[33] Markovski, Yaniv. "How Your Data Is Used to Improve Model Performance." *OpenAI*, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. Accessed 15 May 2023.
[34] C., Johanna. "How Can I Delete My Account?" *OpenAI*, May 2023, https://help.openai.com/en/articles/6378407-how-can-i-delete-my-account. Accessed 15 May 2023.
[35] "New Ways to Manage Your Data in ChatGPT." *OpenAI*, 25 Apr. 2023, https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt.

Children's Safety

One critical focus of our safety efforts is to protect children. We do not permit our technology to be used to generate hateful, harassing, violent or adult content, among other categories,[36] and have designed mitigations to help enforce these policies. GPT-4 is 82% less likely to respond to requests for disallowed content compared to GPT-3.5, and we use a robust combination of human and automated review processes to monitor for misuse. Although these systems are not perfect, we have made significant progress, and are regularly exploring new ways to make our systems safer and more reliable.

We have taken other significant measures to minimize the potential for our models to generate content that may be harmful to children. For example, when users try to upload known Child Sexual Abuse Material to our image tools, we use Thorn's Safer[37] service to detect, review, block, and report the activity to the National Center for Missing and Exploited Children.

In addition to our default safety guardrails, we work with developers such as the non-profit Khan Academy—which has built an AI-powered assistant that functions as both a virtual tutor for students and a classroom assistant for teachers[38]—on tailored safety mitigations. We are also working on features that will allow developers to set stricter standards for model outputs to better support developers and users who want such functionality.

Accuracy

Our models do not answer queries by retrieving or accessing data in a database or on the web;[39] they predict answers based, in large part, on the likelihood of words appearing in connection with one another. In some circumstances, the most likely words that appear near each other may not be the most accurate ones, and the outputs of ChatGPT or other AI tools may also be inaccurate.

Improving factual accuracy is a significant focus for OpenAI and many other AI researchers, and we continue to make progress. We have improved the factual accuracy of GPT-4, which is 40% more likely to produce factual content than GPT-3.5.[40] We also use user feedback on ChatGPT outputs that were flagged as incorrect to improve ChatGPT's accuracy, and since the launch of the product, we have made ChatGPT less likely to generate inaccurate information about people.

---

[36] "Usage Policies." *OpenAI*, 23 Mar. 2023, https://openai.com/policies/usage-policies.
[37] "Our Approach to AI Safety." *OpenAI*, 5 Apr. 2023, https://openai.com/blog/our-approach-to-ai-safety. "Customers." Safer, https://safer.io/customers/. Accessed 15 May 2023.
[38] "Khan Academy." *OpenAI*, 14 Mar. 2023, https://openai.com/customer-stories/khan-academy.
[39] An exception is when the user enables the "browsing" feature on ChatGPT to have the models answer queries by searching the web, rather than answering by itself. In those cases, the answers are based on the model summarizing the web search results.
[40] GPT-4 System Card." *OpenAI*, 23 Mar. 2023, p. 25, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

When users sign up to use ChatGPT, we strive to make it clear that its answers may not always be factually accurate. However, we recognize that there is more work to do to educate users about the limitations of AI tools, and to reduce the likelihood of inaccuracy. Minimizing inaccurate responses is an active research question that we and other AI labs are working on, and we are optimistic about techniques to help address this issue.

Disinformation

OpenAI recognizes the potential for AI tools to contribute to disinformation campaigns. Fighting disinformation takes a whole-of-society approach, and OpenAI has engaged with researchers and industry peers early on to understand how AI might be used to spread disinformation. For example, we recently published work with researchers from Stanford and Georgetown Universities highlighting risks that might arise from disinformation campaigns misusing LLMs, as well as a set of potential policy tools that might help address the issue, such as content provenance standards.[41] As noted above, our Usage Policies also expressly prohibit the use of its tools for certain activities, including generation of violent content, malware, fraudulent activity, high-volume political campaigning, and other areas.[42]

Generating content is only one part of the disinformation lifecycle; false or misleading information also requires distribution to cause significant harm. We will continue to explore partnerships with industry and researchers, as well as with governments, that encompass the full disinformation lifecycle.

Cybersecurity

We understand that our models and tools can have significant impacts on the world, so we dedicate significant resources to maximizing protection of OpenAI's technology, intellectual property, and data.[43] We maintain strict internal security controls and are constantly innovating to improve our defenses. We regularly conduct internal and third-party penetration testing, and audit the suitability and effectiveness of our security controls. We are also building novel security controls to help protect core model intellectual property. We aim to make OpenAI's security program as transparent as possible. Our Trust Portal allows customers and other stakeholders to review our security controls and audit reports.[44]

Other steps we take to improve cybersecurity include the following:

- OpenAI deploys its most powerful AI models as services in part to protect its intellectual property. We do not distribute weights for such models outside of OpenAI and our

[41] Goldstein, Josh A., et al. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *arXiv*, Jan. 2023, https://arxiv.org/pdf/2301.04246.pdf.
[42] "Usage Policies." *OpenAI*, 23 Mar. 2023, https://openai.com/policies/usage-policies.
[43] "Security & Privacy." *OpenAI*, https://openai.com/security. Accessed 15 May 2023.
[44] "Trust Portal." *OpenAI*, https://trust.openai.com/. Accessed 15 May 2023.

technology partner Microsoft, and we provide third-party access via API so the model weights, source code, and other sensitive information stay within OpenAI. OpenAI continuously improves its defenses to prepare for emerging threats.
- We work with partners and cloud providers to protect our models at the data center level.
- OpenAI's security program is built to take into account potential insider threats, and we have built controls to prevent and monitor for model and data exfiltration.
- We recently announced the launch of a bug bounty program inviting independent researchers to report vulnerabilities in our systems in exchange for cash rewards.[45] We also have a dedicated channel for reporting model safety issues, such as the model's response being inaccurate or potentially harmful.[46]

We also recognize that AI tools can be used to both defend against, as well as carry out, cyber attacks. We are committed to evaluating the impact of our models themselves on cybersecurity and working to prevent their misuse. For example, we are in the process of establishing a Cybersecurity Grant Program which will fund researchers conducting security research on salient defensive topics such as training defensive cybersecurity agents, mitigating social engineering, identifying and patching security issues in code, automating incident triage, and other issues. We have established a Cybersecurity Working Group to research how to prevent and protect against AI cyber threats, and our products undergo rigorous testing to limit harms—for example, third-party security researchers were given early access to GPT-4 to test its security capabilities.

Continuing Improvements to Our Safety Approach

We think it is important that our safety approaches are externally validated by independent experts, and that our decisions are informed at least in part by independent safety and risk assessments. For example, in preparing for the GPT-4 release, we facilitated a preliminary model evaluation by the Alignment Research Center (ARC) of GPT-4's ability to carry out certain autonomous actions.[47] We are currently exploring additional possibilities for external validation and testing of our models.

We will also be increasingly cautious with the creation and deployment of more capable models, and will continue to enhance safety precautions as our AI systems evolve.[48] We are investing in developing enhanced evaluations for more powerful models, including assessing AI models for capabilities that could be significantly destabilizing for public safety and national security, so we can develop appropriate mitigations prior to deployment. Addressing safety issues also requires extensive discussion, experimentation, and engagement, including on the bounds of AI system

---

[45] "Announcing OpenAI's Bug Bounty Program." *OpenAI*, 11 Apr. 2023, https://openai.com/blog/bug-bounty-program.
[46] "Model Behavior Feedback." *OpenAI*, https://openai.com/form/model-behavior-feedback. Accessed 15 May 2023.
[47] GPT-4 System Card." *OpenAI*, 23 Mar. 2023, p. 15, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[48] "Planning for AGI and Beyond." *OpenAI*, 24 Feb. 2023, https://openai.com/blog/planning-for-agi-and-beyond.

behavior.[49] We have and will continue to foster collaboration and open dialogue among stakeholders to create a safe AI ecosystem.[50]

Economic Impacts

We understand that new AI tools can have profound impacts on the labor market. As part of our mission, we are working to understand the economic impacts of our products and take steps to minimize any harmful effects for workers and businesses. We are excited to partner closely with leading economists to study these issues, and recently published a preliminary analysis of the economic implications of language models and the software built on top of them.[51] We expect significant economic impacts from AI in the near-term, including a mix of increased productivity for individual users, job creation, job transformation, and job displacement. We are actively seeking to understand the relative proportions of these factors and are eager to work closely with the U.S. government on these issues, including by sharing data and partnering on research.

We believe it is also important to begin work now to prepare for a range of potential scenarios. We are funding research into potential policy tools and support efforts that might help mitigate future economic impacts from technological disruption, such as modernizing unemployment insurance benefits and creating adjustment assistance programs for workers impacted by AI advancements. Our goal is of course not to shape these policies directly, but rather to provide the support and insights policymakers need to understand the potential timeline and extent of impacts of this new technology on the economy. We also support maximizing broad public awareness and understanding of AI technology, particularly through training and education programs for workers in roles and occupations likely to be impacted by AI.

Working with Governments

OpenAI believes that regulation of AI is essential, and we're eager to help policymakers as they determine how to facilitate regulation that balances incentivizing safety while ensuring that people are able to access the technology's benefits. It is also essential that a technology as powerful as AI is developed with democratic values in mind. OpenAI is committed to working with U.S. policymakers to maintain U.S. leadership in key areas of AI and to ensuring that the benefits of AI are available to as many Americans as possible.

We are actively engaging with policymakers around the world to help them understand our tools and discuss regulatory options. For example, we appreciate the work National Institute of Standards and Technology has done on its risk management framework, and are currently researching how to specifically apply it to the type of models we develop. Earlier this month, we

---

[49] "How Should AI Systems Behave, and Who Should Decide?" *OpenAI*, 16 Feb. 2023, https://openai.com/blog/how-should-ai-systems-behave.
[50] "Best Practices for Deploying Language Models." *OpenAI*, 2 June 2022, https://openai.com/blog/best-practices-for-deploying-language-models.
[51] Eloundou, Tyna, et al. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." *arVix*, 27 Mar. 2023, https://arxiv.org/pdf/2303.10130.pdf.

discussed AI with the President, Vice President, and senior White House officials, and we look forward to working with the Administration to announce meaningful steps to help protect against risks while ensuring that the United States continues to benefit from AI and stays in the lead on AI.

To that end, there are several areas I would like to flag where I believe that AI companies and governments can partner productively.

First, it is vital that AI companies–especially those working on the most powerful models–adhere to an appropriate set of safety requirements, including internal and external testing prior to release and publication of evaluation results. To ensure this, the U.S. government should consider a combination of licensing or registration requirements for development and release of AI models above a crucial threshold of capabilities, alongside incentives for full compliance with these requirements.

Second, AI is a complex and rapidly evolving field. It is essential that the safety requirements that AI companies must meet have a governance regime flexible enough to adapt to new technical developments. The U.S. government should consider facilitating multi-stakeholder processes, incorporating input from a broad range of experts and organizations, that can develop and regularly update the appropriate safety standards, evaluation requirements, disclosure practices, and external validation mechanisms for AI systems subject to license or registration.

Third, we are not alone in developing this technology. It will be important for policymakers to consider how to implement licensing regulations on a global scale and ensure international cooperation on AI safety, including examining potential intergovernmental oversight mechanisms and standard-setting.

Conclusion

This is a remarkable time to be working on AI technology. Six months ago, no one had heard of ChatGPT. Now, ChatGPT is a household name, and people are benefiting from it in important ways.

We also understand that people are rightly anxious about AI technology. We take the risks of this technology very seriously and will continue to do so in the future. We believe that government and industry together can manage the risks so that we can all enjoy the tremendous potential.

Biography

Sam Altman: Sam Altman is the co-founder and CEO of OpenAI, the AI research and deployment company behind ChatGPT and DALL·E. Sam was president of the early-stage startup accelerator Y Combinator from 2014 to 2019. In 2015, Sam co-founded OpenAI as a

nonprofit research lab with the mission to build general-purpose artificial intelligence that benefits all humanity. The company remains governed by the nonprofit and its original charter today.