OpenAI

1960 Bryant Street
San Francisco, CA 94110

June 22, 2023

From:
Sam Altman
CEO
OpenAI

To:
Madeline Lubeck
Hearing Clerk
Senate Committee on the Judiciary

Questions for the Record

Responding to Senator Kennedy

1. What are the most important factors for Congress to consider when crafting legislation to regulate artificial intelligence?

Any new laws related to AI will become part of a complex legal and policy landscape. A wide range of existing laws already apply to AI, including to our products. And in sectors like medicine, education, and employment, policy stakeholders have already begun to adapt existing laws to take account of the ways that AI impacts those fields. We look forward to contributing to the development of a balanced approach that addresses the risks from AI while also enabling Americans and people around the world to benefit from this technology.

We strongly support efforts to harmonize the emergent accountability expectations for AI, including the efforts of the NIST AI Risk Management Framework, the U.S.-E.U. Trade and Technology Council, and a range of other global initiatives. While these efforts continue to progress, and even before new laws are fully implemented, we see a role for ourselves and other companies to make voluntary commitments on issues such as pre-deployment testing, content provenance, and trust and safety.

We are already doing significant work on responsible and safe approaches to developing and deploying our models, including through red-teaming and quantitative evaluation of potentially dangerous model capabilities and risks. We report on these efforts primarily through a published document that we currently call a System Card. We are refining these approaches in tandem with the broader public policy discussion.

For future generations of the most highly capable foundation models, which are likely to prove more capable than models that have been previously shown to be safe, we support the development of registration, disclosure, and licensing requirements. Such disclosure could help provide policymakers with the necessary visibility to design effective regulatory solutions, and get ahead of trends at the frontier of AI progress. To be beneficial and not create new risks, it is crucial that any such regimes prioritize the security of the information disclosed. Licensure is common in safety-critical and other high-risk contexts, such as air travel, power generation, drug manufacturing, and banking. Licensees could be

required to perform pre-deployment risk assessments and adopt state-of-the-art security and deployment safeguards.

There remain many open questions in the design of registration and licensing mechanisms for achieving accountability at the frontier of AI development, such as precisely how to define such models, and how to ensure that smaller firms are not unduly burdened with unnecessary regulation. We look forward to collaborating with policymakers in addressing these questions.

2. Can you provide a list of the dangerous capabilities evaluations you use when testing a model?

When we released GPT-4, we published a document that we call a System Card, which described our approach to assessing the potential for dangerous capabilities in GPT-4, gave specific examples of capabilities we investigated, described what we found, and explained what we did to address the identified risks. I am enclosing a copy of the GPT-4 System Card for inclusion in the Senate record. You will find a concise list of safety challenges at page 4 of the System Card, details on our evaluations at pages 4-20, and details on mitigations implemented at pages 21-25.

The development of safety best practices, including standardized evaluations for dangerous capabilities, is an active focus of technical research and engineering effort not only at OpenAI but also at other frontier labs such as Anthropic, Google Deepmind, and Microsoft. We anticipate that these technical efforts will evolve in tandem with public policy work to define meaningful and effective yardsticks for measuring dangerous capabilities and mitigating risks.

3. Can you provide a specific list of tests that a model has to pass prior to being deployed into the world?

We are constantly refining our process for deciding which models to deploy. The description below is a snapshot of our current approach. We intend to share more about our approach publicly in the future, and would be pleased to discuss further with you or your staff at any time.

We use a wide range of evaluations for both the capability and safety dimensions of our models. These include not only quantitative evaluations but also qualitative assessments such as red teaming and engagement with external experts in specific risk areas.

Depending on the novelty and level of capability of a particular model, we apply different evaluation processes. For example, any completely new

model will undergo extensive testing; small tweaks to existing models are subject to a subset of these tests.

We currently distinguish between two different types of review for deployment decisions. Some models undergo heightened and specialized review by a safety board that performs (or instructs employees to perform) further technical and social impact analyses. These are generally models likely to have significantly greater and more novel capabilities than our previously deployed models, often due to significantly superior compute resources or technical parameters employed in training them. Models with technical features comparable to already deployed models are not routed through this review board, though they may still receive a range of evaluations, as described below and in the attached System Card.

OpenAI will sometimes set specific targets for what constitutes sufficiently safe performance on a particular quantitative metric, such as refusing a certain percentage of outputs in a particular risk category, or achieving a sufficient degree of parity in performance across different types of inputs. At other times we use more holistic and qualitative criteria such as the extent to which our understanding of risk (via red teaming, expert engagement, etc.) has "converged" to a reasonably stable point that we believe to be sufficiently safe and our mitigations sufficient to address the identified risks.

Over time, in line with our iterative deployment approach, lessons from early access (e.g. researcher access, red teaming, alpha testing) and larger scale access feed back into improvement of mitigations for those same deployed models, as well as institutional learning for improving the deployment process for future models.

Responding to Senator Tillis

1. What do you see as the principal threats, posed by generative AI, to creators of copyrighted works? How do you plan to address these threats?

We have spoken with and heard from many creators who hold a broad spectrum of opinions about the opportunities as well as the potential risks of AI. Many creators are excited about the opportunities that AI brings to simplify and streamline their work, and to grant them new capabilities in their fields. AI systems can help create initial drafts for consideration, help do preliminary research for written works, or allow creators to quickly iterate on ideas and sketches. For example, DALL-E 2 helped a group of designers and artists create a new cover for an issue of Cosmopolitan Magazine, through an iterative, human-driven process.[1]

[1] https://www.cosmopolitan.com/lifestyle/a40314356/dall-e-2-artificial-intelligence-cover/

We are also aware that many creators are concerned about being displaced or replaced by AI. We've spoken with prominent artists, musicians, writers, actors, filmmakers, and other groups about these issues. Creators are concerned that improperly developed AI systems may directly copy or create derivatives of their work. Many creators also wonder what it will mean to be a human creator in a world where AI can "create" things at significantly greater scale and speed. Some creators are also concerned that the value created by AI systems will accrue to large companies (such as technology platforms or industry rights groups), diminishing the value and power of individual creativity.

Other creators have provided prominent counterpoints to these concerns. They have noted that the role of a human creator is to be able to see and bring to life something that no one else can see, and AI will not replace that human spark, nor will AI be able to "see" such things in the first place. They have also observed that AI is simply the latest in a long line of technological tools that have assisted and empowered creators; the newest generation of creators is already cutting its teeth on these tools, and they will quickly become ubiquitous and essential to creative workflows as many technological tools, such as CGI or auto-tune, are today.

OpenAI does not want to replace creators. We want our systems to be used to empower creativity, and to support and augment the essential humanity of artists and creators. We design and continue to refine our systems to support these goals.

We continue to engage with a broad variety of creators and creative groups about these issues, and are committed to taking part in industry and policy discussion on these topics and to updating our approaches and thinking based on this feedback.

2. For the purposes of determining the presence of copyrighted works in training data and models, to what extent has OpenAI kept records regarding what has been ingested by each iteration of ChatGPT and Jukebox?

The training data for ChatGPT is based on publicly accessible content, licensed content, and data from human trainers. OpenAI keeps records related to these development activities, and has previously disclosed that the vast majority of the publicly accessible content originates from public sources such as Common Crawl and Wikipedia.

The training data for Jukebox was discussed in a paper published about Jukebox at the time of its release in 2020.[2] This model is a

[2] https://arxiv.org/abs/2005.00341

non-commercial research project that was designed to advance scientific progress and understanding of music, and the license attached to the project further stipulates that it may only be used for non-commercial purposes. The open source license is viewable at the Github repository for this project.[3] OpenAI engaged in a dialogue with a leading music industry organization about Jukebox around the time of its release, and has not undertaken any significant work on this model since its release. While there may be many interesting commercial applications in generative AI music that simultaneously respect the work of creators and advance the arts and sciences, it is not a current focus area for OpenAI's activities.

3. Moving forward, are there plans to give creators control over inclusion of their works in training data and how their works are used by a given model? What about works that have already been used in existing training data and models without any creator control?

As noted above and in numerous public statements, OpenAI is committed to engaging with and respecting creators and the creative process. OpenAI's training data comes from a variety of sources, including from publicly available resources like Common Crawl and also by licensing content from content owners.

One method of enabling control is by providing a path for creators to elect not to have a work used to help teach AI models. While we do not believe that as a matter of public policy that anyone's ability to learn from a work should be restricted, we understand concerns from creators about their works being used to help teach AI models and the fear of displacement or replacement. We continue to have cooperative and productive discussions with creators about the use of their content in mutually beneficial and agreeable ways. Resources like Common Crawl respect industry standard means for content owners to disallow use of their content in that resource, so many content owners already have the ability to restrict their works from being used to teach AI models. We are also working on a number of additional ideas in this space. We continue to engage with a broad variety of creators and creative groups about these issues, and are committed to taking part in both industry and public policy discussion on these topics in the future and to updating our approaches and thinking based on this feedback.

Responding to Senator Durbin

1. In February, the Judiciary Committee held a hearing on kids' online safety. During that hearing, witness Emma Lembke testified regarding the toll social media took on her as she grew up. She explained, "As my screen time steadily increased, my mental and physical health suffered."

[3] https://github.com/openai/jukebox

This is an experience shared by too many kids today. From 2015 to 2021, the time kids spent each day on social media rose to nearly three hours—an increase of almost 60 percent in six years. Over a similar time period, CDC data showed a massive spike in negative mental health outcomes for our kids—particularly teen girls. By 2021, 42 percent of teens reported persistent feelings of sadness or hopelessness, and nearly one in three teenage girls said they had seriously considered suicide.

I am concerned that artificial intelligence could exacerbate this problem.

a. In light of our experience with social media, is there a way to safely deploy artificial intelligence so it does not make the current mental health crisis our kids are experiencing even worse?

b. If so, what specific protections are necessary to minimize the potential harms artificial intelligence may pose to kids?

We share the widespread interest in protecting children from negative experiences, and in enabling them to have positive experiences, with new technologies.

We do not sell ads, build user profiles, or attempt to maximize the amount of time that people spend with our products. In many cases, our best experiences involve people spending only a short time with ChatGPT, because it quickly and efficiently does what the person has asked. Our focus is on making our services useful, not addictive.

We employ a neutral age screening mechanism to disallow users under the age of 13, and our legal terms require users ages 13-18 to obtain parental consent. We are also evaluating other improvements to age screening techniques. If we become aware that a child under age 13 has provided personal information directly to OpenAI, we delete that data and do not use it for any purpose. If we determine that an account was created by a child under age 13, we close the account.

We train our systems to refuse to generate hateful, harassing, violent, suicidal, self-harm or adult content. Our Trust & Safety efforts also involve automated and human review processes to monitor for misuse. And if a user tries to upload known Child Sexual Abuse Material to our image tools, we use Thorn's Safer to detect, review, block, and report the activity to NCMEC.

We regularly hear from educators who want to find a way to use our services to help them teach. We are working on a version of ChatGPT for businesses and organizations, including educational institutions, to allow them to share the benefits of this technology with their constituents and

users in a safe and responsible way. We will continue to engage with experts so that any educational products that may be directed to young people in the future are developed with their safety and well-being in mind.

2. What specific guardrails and/or regulations do you support that would allow society to benefit from advances in artificial intelligence while minimizing potential risks?

Any new laws related to AI will become part of a complex legal and policy landscape. A wide range of existing laws already apply to AI, including to our products. And in sectors like medicine, education, and employment, policy stakeholders have already begun to adapt existing laws to take account of the ways that AI impacts those fields. We look forward to contributing to the development of a balanced approach that addresses the risks from AI while also enabling Americans and people around the world to benefit from this technology.

We strongly support efforts to harmonize the emergent accountability expectations for AI, including the efforts of the NIST AI Risk Management Framework, the U.S.-E.U. Trade and Technology Council, and a range of other global initiatives. While these efforts continue to progress, and even before new laws are fully implemented, we see a role for ourselves and other companies to make voluntary commitments on issues such as pre-deployment testing, content provenance, and trust and safety.

We are already doing significant work on responsible and safe approaches to developing and deploying our models, including through red-teaming and quantitative evaluation of potentially dangerous model capabilities and risks. We report on these efforts primarily through a published document that we currently call a System Card. We are refining these approaches in tandem with the broader public policy discussion.

For future generations of the most highly capable foundation models, which are likely to prove more capable than models that have been previously shown to be safe, we support the development of registration, disclosure, and licensing requirements. Such disclosure could help provide policymakers with the necessary visibility to design effective regulatory solutions, and get ahead of trends at the frontier of AI progress. To be beneficial and not create new risks, it is crucial that any such regimes prioritize the security of the information disclosed. Licensure is common in safety-critical and other high-risk contexts, such as air travel, power generation, drug manufacturing, and banking. Licensees could be required to perform pre-deployment risk assessments and adopt state-of-the-art security and deployment safeguards.

There remain many open questions in the design of registration and licensing mechanisms for achieving accountability at the frontier of AI development, such as precisely how to define such models, and how to ensure that smaller firms are not unduly burdened with unnecessary regulation. We look forward to collaborating with policymakers in addressing these questions.

3. During the hearing, you testified that "a new framework" is necessary for imposing liability for harms caused by artificial intelligence—separate from Section 230 of the Communications Decency Act—and offered to "work together" to develop this framework. What features do you consider most important for a liability framework for artificial Intelligence?

Any new framework should apportion responsibility in such a way that AI services, companies who build on AI services, and users themselves appropriately share responsibility for the choices that they each control and can make, and have appropriate incentives to take steps to avoid harm.

OpenAI disallows the use of our models and tools for certain activities and content, as outlined in our usage policies.[4] These policies are designed to prohibit the use of our models and tools in ways that may cause individual or societal harm. We update these policies in response to new risks and updated information about how our models are being used. Access to and use of our models are also subject to OpenAI's Terms of Use which, among other things, prohibit the use of our services to harm people's rights, and prohibit presenting output from our services as being human-generated when it was not.[5]

One important consideration for any liability framework is the level of discretion that should be granted to companies like OpenAI, and people who develop services using these technologies, in determining the level of freedom granted to users. If liability frameworks are overly restrictive, the capabilities that are offered to users could in turn be heavily censored or restricted, leading to potentially stifling outcomes and negative implications for many of the beneficial capabilities of AI, including free speech and education. However, if liability frameworks are too lax, negative externalities may appear where a company benefits from lack of oversight and regulation at the expense of the overall good of society. One of the critical features of any liability framework is to attempt to find and continually refine this balance.

Given these realities, it would be helpful for an assignment of rights and responsibilities related to harms to recognize that the results of AI systems are not solely determined by these systems, but instead respond

[4] https://platform.openai.com/docs/usage-policies/use-case-policy
[5] https://openai.com/policies/terms-of-use

to human-driven commands. For example, a framework should take into account the degree to which each actor in the chain of events that resulted in the harm took deliberate actions, such as whether a developer clearly stipulated allowed/disallowed usages or developed reasonable safeguards, and whether a user disregarded usage rules or acted to overcome such safeguards.

AI services should also be encouraged to ensure a baseline of safety and risk disclosures for our products to minimize potential harm. This thinking underlies our approach of putting our systems through safety training and testing prior to release, frank disclosures of risk and mitigations, and enforcement against misuse. Care should be taken to ensure that liability frameworks do not inadvertently create unintended incentives for AI providers to reduce the scope or visibility of such disclosures.

Furthermore, many of the highest-impact uses of new AI tools are likely to take place in specific sectors that are already covered by sector-specific laws and regulations, such as health, financial services and education. Any new liability regime should take into consideration the extent to which existing frameworks could be applied to AI technologies as an interpretive matter. To the extent new or additional rules are needed, they would need to be harmonized with these existing laws.

Responding to Senator Blumenthal

1. Training data is crucial to foundational models like GPT-4, where content such as news, art, music, and research papers are used to create and refine AI systems, largely material aggregated from the internet. This content represents the labor, livelihoods, and careers of artists, experts, journalists, and scientists.

How should we make sure AI systems respect, acknowledge, and compensate the labor of individuals whose work is used to train AI models?

Ensuring that the creator economy continues to be vibrant is an important priority for OpenAI. Writers, artists, composers and other creators have contributed immeasurably to societies throughout the history of civilization, and they are a vital part of American society and the American economy today. OpenAI is actively engaged in discussions with a wide variety of creators and content owners, geared toward finding mutually beneficial opportunities for creators and technology providers. Those discussions include a recognition by all parties that the technology is still in a nascent stage, and many creators continue to experiment with AI tools to assist in their creation of new works. A few examples:

- Karen Cheng, an artist who uses OpenAI's image generation tool to prompt the AI system to generate creative imagery overlaid to the rhythm of music in the background, created this DALL-E "music video: https://www.youtube.com/watch?v=QM6_YNNwZjk
- Tim Boucher, a science fiction writer, has used a combination of AI tools to write a series of books in a volume driven format that previously would not have been possible: https://www.newsweek.com/ai-books-art-money-artificial-intelligence-1799923
- Paul McCartney is using AI to create a final Beatles album: https://www.npr.org/2023/06/13/1181906529/beatles-john-lennon-voice-song-ai

AI can also be used to make creative works more accessible and available to consumers. For example, OpenAI works with Be My Eyes to provide AI assisted visual description tools that enable visually impaired persons to view and appreciate art they would not have been able to see before.[6]

In addition, OpenAI provides its DALL-E tool in conjunction with an artist financial assistance program, which enables artists to apply for subsidized access to these technologies to ensure they remain broadly available regardless of economic means.[7]

As for enabling artist control, OpenAI's training data comes from a variety of sources, including from publicly available resources like Common Crawl and also by licensing content directly from content owners. Resources like Common Crawl respect industry standard means for content owners to disallow use of their content in that resource, so many content owners have the ability to restrict their works from being used to teach AI models. We are working on a number of ideas in this space.

In addition, OpenAI has had cooperative and productive discussions with creators and creative platforms both about the use of works that could be used to teach AI models, as well as the best ways for such works to be found, showcased, and inspire others in an AI enabled world. OpenAI has also entered into license agreements to pay for specialized content, such as its partnership with Shutterstock, and it expects to continue to do so in the future.

It is also important to build shared understanding about how training data is actually used in development of our models. Some have suggested that generative AI outputs are simply reproductions and rearrangements of pieces of creative expression received via inputs. Others have even claimed that outputs are akin to "collages" or "remixes."

[6] https://openai.com/customer-stories/be-my-eyes
[7] https://openai.com/blog/dall-e-now-available-in-beta

This is not the case. Rather, training data is used to teach models statistical patterns at massive scale. For example, training data for large language models is not used to teach these systems to copy the data, but to learn the fundamentals of language – including vocabulary, grammar, sentence structure, and even basic logic. Ultimately, these AI systems are not search engines or databases, and are not designed to repeat or even store the content on which they are trained. In some ways, the process is akin to a person participating in a foreign language immersion environment and developing an understanding of how to speak, write and think in that new language. The fluency that results is not due to memorizing and parroting specific parts of the thousands of works that may be read or seen or heard. Rather, it is the result of learning and mastering the fundamentals of the new language and applying it, independently, to new situations and ideas. This is the fundamental relationship between AI training data and its outputs.

OpenAI does not want to replace creators. We want our systems to be used to empower creativity, and to support and augment the essential humanity of artists and creators. We are continuously designing and refining our systems. We look forward to continuing our important dialogues with content creators, and with policymakers, to help identify the right frameworks that continue to advance scientific progress while supporting the flourishing of beneficial and valuable creative works.