OpenAI

1960 Bryant Street
San Francisco, CA 94110

March 27, 2024

From:
Anna Makanju
Vice President of Global Affairs
OpenAI

To:
The Honorable Alan Davidson
Administrator
National Telecommunications and Information Administration (NTIA)
1401 Constitution Ave, NW
Washington, D.C. 20230

RE: Request for Comment (RFC) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights (Docket No. NTIA–2023–0009)

OpenAI appreciates the opportunity to provide input on the important questions raised in the NTIA's request for comment on dual-use foundation models with widely available weights.

There are many paths to safe and beneficial AI.

OpenAI believes that building, broadly deploying, and using AI can improve people's lives and unlock a better future. Progress relies on innovation and free market competition. Within those broad guidelines, there are many different paths by which people can further the promise of AI. OpenAI was among the first AI developers to wrestle with the question of how to distribute the benefits of unprecedentedly-capable foundation models, and we start by providing this historical context to help inform the NTIA's deliberations.

In 2019, we created GPT-2, which had the new capability of generating coherent paragraphs of text, and were faced with the question of how to deploy it. On the one hand, the model seemed very useful; on the other hand, we weren't sure if it could be useful for malicious purposes such as phishing email generation. We opted to experiment with a "staged release". As we wrote at the time, "staged release involves the gradual release of a family of models over time. The purpose of our staged release of GPT-2 is to give people time to assess the properties of these models, discuss their societal implications, and evaluate the impacts of release after each stage." When we did not observe significant misuse effects, this gave us the confidence to openly release the full model weights.

In 2020, we created GPT-3, which was much more capable than any previous language model on every benchmark, and again faced the question of how to release it. This time, we decided to release it via our first product, the OpenAI API (an Application Programming Interface, which allows developers to build apps on our technology). As we wrote at the time, we had several motivations for this new release strategy: "commercializing the technology helps us pay for our ongoing AI research, safety, and policy efforts" and "the API model allows us to more easily respond to misuse of the technology. Since it is hard to predict the downstream use cases of our models, it feels inherently safer to release

them via an API and broaden access over time, rather than release an open source model where access cannot be adjusted if it turns out to have harmful applications." Over several years, this API release taught us and the community lessons about the safety and misuse patterns of GPT-3 level models.

In the years since, we have continued to support and believe in the promise of the open-source AI ecosystem, including by openly releasing the weights of some of our state-of-the-art models (such as CLIP and Whisper) and developing open-source infrastructure for other AI developers (such as the Triton GPU programming language). We have seen openly released weights bring a variety of significant benefits, including facilitating academic research on the internals of AI models, enabling users and organizations to run models locally on their edge devices, and facilitating creative modifications of models to suit users' ends. Many AI companies have chosen to invest heavily in open model weight releases for a variety of reasons, including brand, recruiting, and attracting a developer ecosystem to build on and accelerate the internals of a company's technology.

At the same time, our approach to releasing our flagship AI models via APIs and commercial products like ChatGPT has enabled us to continue studying and mitigating risks that we discovered after initial release, often in ways that would not have been possible had the weights themselves been released. For example, we recently partnered with Microsoft to detect, study, and disrupt the operations of a number of nation-state cyber threat actors who were abusing our GPT-3.5-Turbo and GPT-4 models to assist in cyberoffensive operations. Disrupting these threat actors would not have been possible if the weights of these at-the-time frontier models had been released widely, as the same cyber threat actors could have hosted the model on their own hardware, never interacting with the original developer. This approach has enabled us to continue to distribute the benefits of AI broadly, including via widely-available free and low-cost services.

These experiences have convinced us that both open weights releases and API and product-based releases are tools for achieving beneficial AI, and we believe the best American AI ecosystem will include both.

Combining iterative deployment with a Preparedness Framework

Over and over again, across both product releases and weight releases, we have seen the incredible benefits of "iterative deployment": gradually putting increasingly capable AI into people's hands so they can use it to improve their lives, and helping society adjust to these new technologies. As we wrote in 2023: "We work hard to prevent foreseeable risks before

deployment, however, there is a limit to what we can learn in a lab. Despite extensive research and testing, we cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it. That's why we believe that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems over time."

As AI models become even more powerful and the benefits and risks of their deployment or release become greater, it is also important that we be increasingly sophisticated in deciding whether and how to deploy a model. This is particularly true if AI capabilities come to have significant implications for public safety or national security. The future presence of such "catastrophic" risks from more advanced AI systems is inherently uncertain, and there is scholarly disagreement on how likely and how soon such risks will arise. We do not believe there is yet sufficient evidence; we can't rule them out, nor be certain they're imminent. As developers advancing the frontier of AI capabilities to maximize their benefits, we view building the science of the risks of this technology (including gathering evidence related to those risks) as integral to our work.

To navigate these uncertainties in an empirically driven manner, OpenAI publicly launched our Preparedness Framework, a science-based approach to continuously assess and mitigate any catastrophic risks that might be posed by our AI models. The Preparedness Framework defines how we evaluate our AI models' capability levels in several high-risk domains, including cybersecurity, autonomous operation, individualized persuasion, and CBRN (Chemical, Biological, Radiological, and Nuclear) threats. For an example of this framework in action, see our recent study testing GPT-4's ability to aid in biological threat creation, which concluded that it poses no significant marginal risk.

Based on these evaluations, we assess models' risk levels in each category as Low, Medium, High, or Critical. Crucially, under our Preparedness Framework, we will not deploy AI systems that pose a risk level of "High" or "Critical" in our taxonomy (and will not even train "Critical" ones, given their level of risk), unless our mitigations can bring these systems' risk down to at most a "Medium" level. The Preparedness Framework is important because it lets us build and widely share the benefits of increasingly capable AI, while preparing us to detect and protect against catastrophic risks as early as possible if they do arise.

Practices for developers of highly capable AI

We believe that people and companies should be able to participate in AI as they choose —which can include developing or using AI that reflects

their values and vision —in order to achieve the benefits of AI. At the same time, highly capable AI systems should be built and used safely, with any discovered catastrophic risks appropriately mitigated. These interests may sometimes be in tension, and need to be thoughtfully managed in a case-specific way to achieve the best outcomes for society.

In the case of highly-capable foundation models which require significant resources to create (on the order of hundreds of millions of dollars or more), we believe that AI developers should assess their model's potential to pose catastrophic risks, and, if the model's risk level is found to be high, put appropriate mitigations in place before deploying or releasing it. This strikes an appropriate balance between risk-management and innovation: these models are anticipated to have the greatest capabilities, while the cost of assessment is at most a small fraction of their development cost. Such assessments make sense regardless of whether the model's weights are intended to be released widely or through an API.

On the other end of the spectrum, in the case of less resource-intensive foundation models, the balance of interests is different. On current evidence, such models appear much less likely to pose catastrophic risks, even with likely advances in finetuning and model-modification techniques. Meanwhile, assessments for catastrophic risk may cost a substantial fraction of the budget of small training runs, which could lead to a chilling effect on innovation and competition. We believe that such assessments for catastrophic risks should not be expected for these models, as there is huge value in protecting a diversity of developers' ability to innovate on exciting new AI capabilities and allowing the marketplace of ideas and products to flourish, and the science indicates that these models' risk is relatively low.

Assessment protocols like the Preparedness Framework are a useful tool to evaluate the *ex ante* risks from any type of model release, including open model weight releases. There are a few considerations that are specific for how to apply them to open weights releases.

One such consideration is that testing conditions would ideally reflect the range of ways that downstream actors can modify the model. One of the most useful properties of open models is that downstream actors can modify the models to expand their initial capabilities and tailor them to the developer's specific applications. However, this also means that malicious parties could potentially enhance the model's harmful capabilities. Rigorously assessing an open-weights release's risks should thus include testing for a reasonable range of ways a malicious party could feasibly modify the model, including by finetuning. OpenAI already conducts some

modification-testing as part of our Preparedness Framework (as we did in our biorisk assessment).

Another key consideration is that open model developers may be unable to rely on system-level safeguards to reduce the risk of their model's misuse, as safeguards can often be removed by a malicious downstream user who possesses the model weights. Today, this difference in mitigation ability has limited consequences, since even our most capable current models are not rated as especially risky. But if a future model is scientifically determined to pose severe risks if released, then the path to reduce the risk of an open-weights release may rely on increasing the resilience of the external environment into which the model is released.

The need for societal resilience to AI misuse is broader than any one organization's release decisions. Given continuing progress in and dissemination of AI algorithms, and increasingly widespread access to compute (including in countries of concern to the United States), today's frontier AI capabilities — often accessible to only a few actors at time of creation — will eventually proliferate widely. The United States, and countries around the world, also have an opportunity to invest in and lead in mitigations that will limit the consequence of misuse, so the balance of outcomes is maximally positive.

For instance, strengthening resilience against AI-accelerated cyberattack risks might involve providing critical infrastructure providers early access to those same AI models, so they can be used to improve cyber-defense (as in the early projects we have funded as part of the OpenAI Cybersecurity Grant Program). Strengthening resilience against AI-accelerated biological threat creation risks may involve solutions totally unrelated to AI, such as improving nucleic acid synthesis screening mechanisms (as called for in Executive Order 14110), or improving public health systems' ability to screen for and identify new pathogen outbreaks. If an AI model is rigorously shown to pose severe risks to public safety or national security, then the developer may also have an important role to play in building awareness of the new capabilities in advance of wide release (such as via notifying infrastructure providers or limiting API deployment), to create both time and motivation for urgently needed resilience efforts. This mirrors the norm of "responsible disclosure" from the cyber domain, where security researchers will temporarily embargo the release of vulnerabilities they find so as to give time for defenders to patch their systems, while not slowing down further security research.

We need a better science of AI risks

While we believe that assessing the risks of the most capable models is important, the science of AI risk evaluations is nascent. OpenAI and the

broader AI community are still building the foundations of how to assess AI risks, and we are still constantly iterating on many of the operationalization details in the Preparedness Framework. Governments have an important role to play in helping the AI ecosystem mature its risk and capability evaluation practices, such as by convening experts from the offensive cybersecurity, critical infrastructure, and AI worlds to agree on a set of priority AI cyber threat models, and build out rigorous and empirical testbeds for assessing them. We strongly support the voluntary innovation-friendly and science-first approach being pursued by the USAISI.

Ever since OpenAI faced the choice of how to release GPT-2 in 2019 – opting to release only a small version of the model at first — new findings and events have continuously changed the landscape of considerations around open release of foundation model weights, sometimes every few months. We expect this trend to continue. Any government policy approach should be flexible and adaptable to future changes.