OpenAI

1960 Bryant Street
San Francisco, CA 94110

February 2, 2024

From:
Anna Makanju
VP of Global Affairs
OpenAI

To:
National Institute of Standards and Technology (NIST)

To whom it may concern:

OpenAI was created as a nonprofit in 2015 to ensure that artificial general intelligence—in short, AI that's at least as smart as a person—benefits all of humanity. We research, develop, and release cutting-edge AI technology as well as tools and best practices for the safety, alignment, and governance of AI. We welcome this opportunity to inform NIST's ongoing and critical work on AI.

Here, we focus on three topics raised in the RFI: (1) evaluating and auditing AI capabilities, (2) conducting red teaming tests to enable deployment of safe, secure, and trustworthy systems, and (3) synthetic media and provenance.

Evaluating dangerous capabilities in AI systems

We applaud NIST's focus on "creating guidance and benchmarks for evaluating capabilities… through which AI could cause harm." OpenAI has committed to a Preparedness Framework, a comprehensive approach to evaluate, track, and mitigate catastrophically dangerous risks from current and future AI models. The Preparedness Framework currently tracks four initial areas of risk: cybersecurity; chemical, biological, nuclear, and radiological threats (CBRN); persuasion; and model autonomy. The Framework also commits us to ongoing vigilance for "unknown unknown" risks that have not yet been identified.

As part of this work, OpenAI recently shared one large-scale evaluation for CBRN: assessing GPT-4's ability to meaningfully increase malicious actors' access to dangerous information about biological threat creation, compared to the baseline of existing resources (i.e., the internet). In the largest-of-its-kind evaluation involving both biology experts and students, we found that GPT-4 provides at most a mild uplift in biological threat creation

information. While not a large enough uplift to be conclusive, we hope this finding serves as a starting point for continued research and community deliberation, which we hope will be driven by NIST and the new AI Safety Institute.

This work increased our confidence in several key principles for evaluating risks from AI systems:

- **AI systems' contribution to risks should be measured in terms of *change* relative to an appropriate baseline.** Many of the risks that may be increased by current and future AI systems (such as in cybersecurity or biosecurity) exist at some level even without AI. For example, internet search already enables a substantial degree of access to biosecurity-relevant information. When evaluating AI systems' contribution to risks, an important best practice is to test whether AI increases risk *beyond* existing resources. In our recent study on biorisks, we operationalized this by randomly assigning half of the participants into a control group that was free to use only non-AI sources of knowledge (including online databases, articles and internet search engines, as well as any of their prior knowledge), and assigning the other half into a treatment group with full access to both these resources and the GPT-4 model.
- **Working with domain experts is vital to understanding risks.** It is challenging for any one entity to hire world-class experts in all of the wide and varied topics that are relevant to AI safety. To access gold-standard expertise, it is useful to partner with third parties that employ domain experts in the subjects relevant to dangerous capabilities evaluations. In addition, involving domain experts in the grading of the studies helps provide assurance that the evaluations are being conducted objectively. For example, in developing and administering the biorisk evaluation, we worked closely with third-party biosecurity experts on designing the research tasks, administering safety trainings for participants, and grading the completed tasks. It would be in the interest of AI safety to expand and diversify this ecosystem.

- **Thorough evaluation also requires working with AI experts to effectively elicit the full range of model capabilities.** To understand the full range of risks from AI models, it's necessary to elicit the full capabilities of the model wherever possible in the evaluation. This requires a deep understanding of the underlying AI systems and how they can be effectively leveraged. We recommend that evaluations be designed in close cooperation with AI experts. In our biorisk study, this included providing training to human subjects on how to get better performance from language model capability elicitation best practices, as well as custom technical approaches to better elicit and probe the capabilities of the models.
- **We need more research on how to interpret results of risk evaluations.** For example, in the case of evaluating AI models' increasing access to biorisk information, it is not yet clear what level of increased information access would translate to significantly increased biorisk. The effect of AI systems on biorisk may change as new technologies emerge that can translate online information into physical biothreats. As we continue to operationalize our Preparedness Framework, we are eager to work with NIST and the AI Safety Institute to build a stronger understanding of risks and risk metrics.
- **Gold-standard human subject evaluations are expensive.** Conducting human evaluations of language models requires a considerable budget for compensating participants, developing software, and security. In our biorisk study, we explored various ways to reduce these costs, but most of these expenses were necessitated by either (1) non-negotiable security considerations, or (2) the number of participants required and the amount of time each participant needs to spend for a thorough examination. This should be taken into account when designing standards.

Additional information is available in our blog post on the recent biorisk study: [Building an early warning system for LLM-aided biological threat creation].

Red teaming to enable deployment of safe AI systems

What is red teaming?
OpenAI defines red teaming as "a structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats."[1] There are various possible methods emerging under the umbrella term of red teaming, including internal red teaming (done by internal, dedicated teams at a lab or company), external red teaming (done by external stakeholders in collaboration with a lab or company), or automated red teaming (using AI models to generate automated attacks and classifying outputs). In the context of this document, we are primarily referring to external red teaming efforts which involve OpenAI working with external domain experts to assess the capabilities and risks of an AI model or system.

OpenAI's approach to red teaming does not consider adversarial attacks or model outputs in isolation. Rather, it is a method for eliciting risks in a contextualized, holistic manner in collaboration with domain experts.[2] In addition to malicious use and methods to circumvent safety mitigations, red teaming also considers other risks: benign or expected inputs leading to harmful or risky outputs, novel capabilities improvements that may alter the risk landscape, and how factors outside of the system itself may interact with model outputs to cause risks or harms. Assessments of these areas often benefit from having humans in the loop to generate potential examples, and to validate the resulting outputs in the context of a given red teamer's expertise.

What is red teaming useful for?
AI red teaming helps to understand the potential risks associated with new models and systems that:

- Require forms of interactions which may be different from previous AI systems or technologies and are not well

[1] See the Frontier Model Forum's definition of red teaming.
[2] We use the term "expert" to refer to expertise informed by a range of domain knowledge and lived experiences.

covered by programmatic evaluations (e.g., inpainting on DALL·E, GPTs).
- Have significantly improved capabilities which may introduce novel risks that have not yet been evaluated (e.g., scientific domains, persuasion, or reasoning).
- Require context or domain specific knowledge for testing and verification (e.g., region-specific political content, cultural biases, scientific or expert domains such as law and medicine).
- Require an understanding of a user flow or specific use cases, including factors that may be external to the system itself (e.g., testing GPT-4(V) for low vision individuals).

OpenAI views red teaming as a tool for assessing both model-level and system-level risks. System features may include: classifiers, prompt filters / block lists, user interface level interventions, monitoring and evaluation practices and other policy enforcement mechanisms. We sometimes conduct red teaming for a new product even when there is not a new model involved. For example, while GPTs did not introduce a new underlying model, they did introduce new systems for how users interact with the model.

OpenAI views our red teaming efforts as complementary to further domain specific red teaming efforts that should be conducted by developers building on top of our technology. For example, while we subject our models and systems to red teaming at specific points in time under particular conditions, developers who are building upon our API should take into consideration those learnings, and conduct additional red teaming based on the system and contextual conditions they expect to operate in. This is one of the reasons why OpenAI publishes the key findings from red teaming efforts in System Cards (and other forms of publicly available documentation) for others to learn from and build upon.

Iterative red teaming at OpenAI
We have documented several of our red teaming efforts for frontier model launches in System Cards:

- DALL·E 2 System Card

- GPT-4 System Card
- GPT-4(V) System Card
- DALL·E 3 System Card

OpenAI has provided expert red teamers access to pre-trained models with varying degrees of fine-tuning and post-training as well as varying maturity levels of safety mitigations.

The goals of doing so are as follows:

- Red teaming insights may inform the development of post-training level mitigations, system level mitigations, policies, and evaluations.
- Red teaming insights may help inform leadership decision-making on releasing certain features, how to iteratively deploy the release, and the effectiveness of safety mitigations.
- Red teaming results may be shared alongside public launch materials (such as in System Cards or other formats) to inform potential users and other stakeholders about risks that have been mitigated, residual risks, and possible future risks.

We engage red teamers as early as is reasonable in the development process, so that red teaming insights can directly feed into safety efforts and decision-making. It is also important to learn about the model's base capabilities prior to any added safety mitigations, so that model developers can make informed decisions about the model's base level risks, and for societal understanding about the risk landscape associated with increasingly powerful systems.

Once safety mitigations have been put into place, red teaming efforts may focus additional rounds of red-teaming on identifying gaps and residual risks that are not addressed by the safety mitigations, as well as assessing the robustness of the mitigations.

Ultimately, while there are important safety properties to consider further upstream of model development processes, red teaming

intends to simulate an experience as close as possible to what model developers release to the public.

Limitations of red teaming
Red teaming in and of itself is not a sufficient risk measurement exercise. On its own, red teaming will not quantify the probability or propensity of a model to produce harmful content or risks associated with the use of an AI system. Red teaming also does not provide enough information to quantify the severity of an identified risk or harm.

While most of OpenAI's expert red teaming efforts take place prior to a major model or product deployment, models and systems evolve quite often in production, and as such, it is important to take that into account when contextualizing red teaming findings. Similarly, developers building for particular use cases on models may make design decisions that alter the safety profile of a model or system if it is not inherent to (or immutable from) the model or system itself.

Red teaming lays the foundation for types of further testing and evaluation, and provides some guidance about attack vectors or issues that safety mitigations need to be robust against.

Examining multiple examples and permutations of an issue can help to instill confidence in how to measure a particular risk area. Expert red teaming by design aims to cover breadth instead of depth of risk areas, and as such, on its own would not necessarily create an evaluation sufficient for measuring specific risks. Instead, red teaming can generate datasets that might be considered the "seeds" for a more thorough evaluation. From there, the results can be used to generate more examples of a particular issue area that was uncovered, and a "golden set" of labeled examples (usually, by domain experts) can be used for evaluating future models on an identified issue area.

Composition of red teams and domain prioritization
General purpose AI systems that will be used for many anticipated and unanticipated use cases and in a variety of contexts around

the world necessitate covering a wide range of topic areas, with people representing a wide range of perspectives and worldviews.

OpenAI believes in recruiting a wide variety of experts to red team our models. Last year, we put out a call for applicants to the Red Teaming Network. Selection criteria included:

- Demonstrated expertise or experience in a particular domain relevant to red teaming
- Passionate about improving AI safety
- Not having any conflicts of interest
- Diverse backgrounds and traditionally underrepresented groups
- Diverse geographic representation
- Fluency in more than one language
- Technical ability (helpful but not required)

Domain prioritization can be informed by: expected uses of the AI systems or model, especially in contexts with higher ambiguity or possible risks, early evaluation of models where model developers might expect increased capabilities, known previous content policy issue areas, and relevant socio-political contexts (e.g, 2024 is a major election year in many places around the world). It is important to note that each model or system may require varying sets of expertise, and new domains may be considered based on the advancing capabilities and novel use cases of model or systems. As such the optimal composition of red teams will evolve over time.

Synthetic Media and Provenance

There are three broad categories of provenance techniques for AI-generated media:

- **Watermarking:** Under this approach, the audiovisual generated media itself contains a signal of its origins – a subtle pattern not apparent to the viewer or listener, but that can be detected by software. This might be a signal that can be detected only with the help of a secret key, or alternatively, the software for detecting the watermark might be publicly available. Because of this, if OpenAI were to add

a watermark to our outputs, collaboration across the AI value chain would be necessary so that other participants, such as social media platforms that distribute content, could make the watermark apparent and useful to users. If the detection process is not itself public, then access to that process is a complex policy question.
There are also technical challenges. Although watermarks may be harder to remove than other provenance methods, the marked media may still lose its watermark if it is cropped, resized or otherwise modified. For these reasons, watermarks can still be evaded, particularly by motivated adversarial actors. In addition, the impact of watermarking may be limited given that bad actors can access models that do *not* watermark their outputs.

- **Classifiers** (trained models that distinguish AI-generated output from other media, and may detect which model or service generated a given output): When they are effective, these approaches are highly appealing because they do not rely on cooperation from the person distributing an image or from anyone else. However, they can make mistakes – both false positives and false negatives – and may be computationally intensive to deploy at scale. False positives might, for instance, incorrectly describe a human artist's work as an AI output. False negatives, on the other hand, may incorrectly flag an image as non-AI generated, when in fact it is.

  Additionally, classifiers may not be as broadly available as other forms of provenance data are. As with watermarking detectors, the question of who gets access to a classifier raises complex policy equities.

- **Metadata-based approaches** (such as C2PA's current standard): In these approaches, the metadata accompanying certain media is cryptographically signed to provide an attestation of the media's origins.

This can empower people who want to prove the origin of media, whether AI-generated or not. For example, C2PA could allow a news publisher to demonstrate, and viewers to confirm, that the publisher actually published a certain image or video and stands by the accuracy of that image or video. Similarly, if implemented for a generative AI system, this technique could help an artist to show that they generated a certain synthetic image or video. These approaches have the benefit, ostensibly, of providing consumer or public visibility into content provenance. Additionally, they have the advantage of not requiring significant resources to implement.

However, metadata can easily be removed from an underlying image or video, so this technique does not create a meaningful barrier for bad actors (for instance those engaged in disinformation campaigns) who might want to pass off generated content as real.

In order for metadata approaches to broadly benefit the public, browsers and distribution platforms, such as social media platforms would need to detect and display the metadata. Successfully implementing metadata-based approaches therefore, requires collaboration throughout the value chain: It is not enough for audiovisual materials to have metadata cryptographically signed, but the distribution platforms must be able to detect the metadata in question and display it for an end-user to verify the media's origins.

OpenAI's approaches to provenance

Because each provenance method comes with advantages and limitations, OpenAI has been exploring a range of approaches to provenance for AI-generated audiovisual media.

C2PA metadata for DALL·E 3 images

On January 15 of this year, OpenAI announced that we will be implementing the C2PA metadata approach for images generated using our text-to-image model DALL·E 3. C2PA specifications are

an open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media.

These specifications allow metadata to be attached to a file. This metadata includes information about the source of an image (in our case, that the image came from DALL·E) and the time of creation. Members of the public can test for this metadata and, if the metadata is present, confirm that an image was generated by DALL·E 3.

This will help us empower users to indicate the origin of images that they generate using DALL·E 3. However, this metadata can be removed fairly easily: a motivated bad actor can remove the C2PA metadata that accompanies any image. In addition, common image sharing platforms like social media platforms currently remove it by default, rather than detecting and presenting it to users. Given how easily C2PA can be removed, members of the public cannot assume that every DALL·E image they see will necessarily have such data.

However, C2PA isn't just for AI images, and it could have important benefits if more broadly adopted. It is also being adopted by camera manufacturers, news organizations, and others to vouch for where images come from. We believe broader adoption of disclosure methods, and encouraging users to look for these signals, are important steps towards increasing trustworthiness of digital information.

Experimental classifier for DALL·E 3 images

On October 19, 2023, we announced our ongoing work on a provenance classifier, a new internal tool for detecting images generated by our DALL·E 3 system. We measure the classifier's accuracy using internal benchmarks which have shown promising results, even where images have been subject to common types of modifications, such as cropping, resizing, JPEG compression, or when text or cutouts from real images are superimposed onto small portions of the generated image. Despite these strong results on internal testing, the classifier can only tell us that an image was

likely generated by DALL·E, and does not yet enable us to make definitive conclusions.

We are continuing to test our classifier for robustness and, in the first quarter of 2024, we plan to make it available to external partners for feedback. In the year ahead, we look forward to beginning to broaden our experiments with the image classifier, by inviting select external parties to join us in assessing its performance and usefulness.

The classifier is tailored to the model and is only able to classify whether an image was likely generated by DALL·E, and therefore, even if it were completely accurate in its classifications, it would not be possible to use it to determine if an image was generated by another generative tool.

We welcome the opportunity to collaborate with you as your work in this area continues.

Sincerely,

Anna Adeola Makanju
VP of Global Affairs
OpenAI