OpenAI

Statement of Aleksander Mądry, Head of Preparedness, OpenAI

Senate AI Insight Forum - December 6, 2023

We are living in remarkable times. Times when we can feel incredibly optimistic about the potential benefits that AI can bring, but also times when we need to recognize the increasing risks these systems may pose. At OpenAI, we want to think deeply and holistically about risks across the spectrum, from the more near-term risks to the longer-term ones. Today, I would like to share a bit about our approach to AI safety, with a specific focus on catastrophic risk, which we are driving with our Preparedness team.

The concept of iterative deployment has driven much of our understanding of AI safety. Iterative deployment, or careful deployment of successively more powerful AI systems, enables us to learn about safety implications based on observed risk and misuse. By informing and deepening our real-world understanding of AI safety, this helps us build technical and procedural mitigations for emerging risks. This gradual process also allows society to better understand and prepare for AI, as the technology evolves.

The importance of empirical understanding is also the core driver of our safety work on catastrophic risks. I would like to share three of the guiding principles that underpin this work.

First, we aim to be driven by facts and science. To this end, we are investing in the design and execution of rigorous capability evaluations and forecasting, so that we can better detect emerging risks. In particular, we want to move the discussions about catastrophic risk beyond hypothetical scenarios to measurements and data-driven predictions of catastrophic risk.

Second, we aim to be proactive about the mitigation of the identified risk. This involves developing technical solutions to not only address current frontier system risk but also to target the systems we will build in the longer-term future. Beyond such technical work, we also aim to improve on our processes and procedures to make sure we can make all this knowledge actionable and to ensure that safety thinking is in OpenAI's DNA.

Third, we aim to holistically reflect the interests of humanity, which requires thinking about benefits in conjunction with risks. As we have said before, we believe artificial general intelligence (AGI) has the potential to increase abundance and serve as a great force multiplier for human innovation, from education to medicine and beyond. It is thus imperative that we take catastrophic risk extremely seriously so as to prepare and protect ourselves from harm, and make sure we realize these upsides of AGI.

As AI development across the industry accelerates, we want to double down on our safety efforts related to such catastrophic risk. To this end, we launched a dedicated team—the Preparedness team, which I have the pleasure of leading—that focuses on research,

evaluations, monitoring, and forecasting of catastrophic risk posed by increasingly powerful models, as well as collaborating on mitigations. The initial focus of this work has been across four priority catastrophic risk areas: individualized persuasion, cybersecurity, CBRN (chemical, biological, radiological, and nuclear) threats, and model autonomy, along with a dedicated workstream for tracking “unknown unknowns”, i.e., risks that we might not yet realize we should be tracking.

In conjunction with building the Preparedness team, we are currently developing our Preparedness Framework, a document describing OpenAI’s processes to identify, track and protect against catastrophic risks. This framework distills our latest learnings on how to best achieve safe development and deployment in practice. It also lays out the processes that will help us rapidly refine our understanding of the science and empirical texture of catastrophic risk, and to mitigate unsafe development.

Our Preparedness Framework involves:

- *Tracking catastrophic risk level via evaluations:* We will be building and continuously improving suites of evaluations and other monitoring solutions across all tracked risk categories. Importantly, we will also be forecasting the future progression of these risks, so that we can develop safety and security mitigations ahead of time.
- *Seeking out unknown-unknowns:* We will continuously run a process for identification and analysis (as well as tracking) of currently unknown categories of catastrophic risk so that we can rapidly detect new risks as they emerge.
- *Establishing safety baselines:* We will establish procedural commitments and thresholds that trigger changes to our security, development, and deployment processes.
- *Driving operationalization via the Preparedness team*: The Preparedness team will drive the technical work and maintenance of the Preparedness Framework. This includes conducting research, evaluations, monitoring, and forecasting of risks, and synthesizing all findings via regular reports. These reports will include a summary of the latest evidence and make recommendations on changes needed to enable OpenAI to plan ahead.
- *Creating a cross-functional advisory body*: We are creating an internal advisory body that brings together expertise from across the company to inform safety decisions.

While work on catastrophic risk from AI has never been more important, these efforts are meant to be just one piece of our overall approach to safety and alignment. We are also working on mitigating bias, hallucination, and misuse, facilitating democratic inputs to AI, improving alignment methods, red-teaming our models, releasing system cards, and investing significantly in security and safety research. This is also one more way in which we are meeting our voluntary commitments to safety, security and trust in AI that we made in July 2023.

OpenAI’s primary fiduciary duty is to humanity. We are committed to doing the research required to make AGI safe and investing in the science and governance to help ensure AI benefits all of humanity.