OpenAI

OpenAI White Paper on the European Union's Artificial Intelligence Act

Introduction

OpenAI is an AI research and deployment company with the mission of ensuring that artificial general intelligence (AGI) is developed and used in a way that benefits all of humanity.[1] Since our founding in 2015, we have deployed numerous AI systems on the path towards that goal, including GPT-3, a large language model that performs a variety of natural language tasks; DALL·E, an image generation system that draws detailed pictures from text input; and Codex, a code generation system which writes code based on text input.

OpenAI's foundational charter revolves around the development of "safe and beneficial AGI." In addition to core AI research and development, we invest heavily in policy research and formulation, risk analysis and mitigation, and technical and process infrastructure to maximize safe use of our technologies. Our company is governed by a non-profit with independent directors making up a majority of the board, and the board is required to put social benefit ahead of all other considerations. OpenAI also has a unique "capped-profit" legal structure that allows us to effectively increase our investments in computing power and talent while maintaining the checks and balances needed to actualize our mission.

We believe that AGI has the potential to profoundly benefit society, and understand that realizing these benefits requires oversight and governance of AI beyond industry alone. We support thoughtful regulatory and policy approaches designed to ensure that powerful AI tools benefit the largest number of people, and we applaud the EU for tackling the immense challenge of comprehensive AI legislation via the Artificial Intelligence Act (AIA).

OpenAI shares the EU's goal of increasing public trust in AI tools by ensuring that they are built, deployed, and used safely, and we believe the AIA will be a key mechanism in securing that outcome. Many themes and requirements of the AIA are reflected in the tools and mechanisms that OpenAI already employs to balance technological progress with safe and beneficial use. For example, we currently require applications building with our tools to adhere to use-case policies that exclude harmful or especially-risky uses; monitor and audit applications to help prevent misuse; and employ an iterative deployment process, through which we release products with baseline capabilities and stringent restrictions, and slowly expand features and/or

[1] We define "artificial general intelligence" as highly autonomous systems that outperform humans at most economically valuable work. More information is available at https://openai.com/charter/.

loosen requirements as we receive feedback on how they are being used. We seek to share our experiences in building and deploying AI systems while continuing to learn from others.

We recognise that the EU has received feedback on all aspects of the AIA - accordingly, this White Paper focuses on issues that we are particularly familiar with given our experience and mission: requirements around general purpose AI systems, modifications to deployed systems, and the scope of certain high risk use cases. We appreciate the opportunity to contribute to the discussion and are excited for ongoing engagement on the AIA.

DISCUSSION

I. General Purpose AI Systems as High-Risk Systems

Recent amendments to the AIA have sought to ensure that general purpose AI systems, which have the potential to be deployed in high risk use cases, are adequately covered by the AIA. In their consolidated proposal (15/06/2022), the French Presidency of the Council has put forward Articles that would cover or exclude general purpose AI systems under certain conditions. We understand the concerns arising from the unregulated release of general purpose AI systems and offer suggestions on potential impact and issues to consider.

For background, OpenAI primarily deploys general purpose AI systems - for example, our GPT-3 language model can be used for a wide variety of use cases involving language, such as summarization, classification, questions and answers, and translation. By itself, GPT-3 is not a high-risk system, but possesses capabilities that can potentially be employed in high risk use cases. Accordingly, we have dedicated significant resources to determining guidelines, best practices, and limitations for uses for our services. We currently outline a set of “high stakes applications” in fields such as law, medicine, politics, finance, and civil services, where applications proposed to be built using our services are subject to additional scrutiny that requires clear identification and management of risks. For example, in an employment context, we would not support a use case involving the use of GPT-3 to determine eligibility for employment, but may support a use case where GPT-3 assists a user by suggesting potential text for job postings (which is reviewed by the user before publication), given the simpler bounds and comparatively lower risk of the latter. This level of oversight over the use of our services is enabled by deploying GPT-3 through an application programming interface (API) which allows us to review signups, implement technical oversight, and identify and prevent repeated acts of abuse.

We believe our approach to mitigating risks arising from the general purpose nature of our systems is industry-leading, and we have outlined some of these practices in a collaborative publication with other labs titled “Best Practices for Deploying Large Language Models.” Despite

measures such as those previously outlined, we are concerned that proposed language around general purpose systems may inadvertently result in all our general purpose AI systems being captured by default.

The currently proposed Article 4.c.1 contemplates that providers of general purpose AI systems will be exempted "when the provider has explicitly excluded any high-risk uses in the instructions of use or information accompanying the general purpose AI system." While we believe that we would currently fall under this exemption given the protective measures we employ, Article 4.c.2 potentially undermines the intent of Article 4.c.1 by stating that "Such exclusion shall…not be deemed justified if the provider has sufficient reasons to consider that the system may be misused."

As outlined above, we consider and continue to review on an ongoing basis the different ways that our systems may be misused, and we employ many protective measures designed to avoid and counter such misuse. The current framing may inadvertently incentivise an avoidance of active consideration of ways that a general purpose AI system may be misused so that providers do not have "sufficient reasons to consider [misuse]" and can avoid additional requirements. The fundamental nature and value of general purpose AI systems are that they can be used for many application areas; we do not think it would meet the goals of safe and beneficial AI to inadvertently encourage providers to turn a blind eye to potential risks.

We suggest reframing the language to incentivize rather than penalize providers that consider and address system misuse, especially if they take actions that indicate they are actively identifying and mitigating risks.

An example of possible language could be that providers of general purpose systems will be exempted as per Article 4.c.1 *"when the provider (i) has explicitly excluded any high-risk uses in the instructions of use or information accompanying the general purpose AI system, (ii) performs periodic assessments to understand the possibility of misuse, and (iii) implements reasonable mitigation measures to address those risks."* We propose removing the language currently in Article 4.c.2 and replacing it with this suggested text.

II. Generative AI systems considered high-risk under the IMCO-LIBE report

The European Parliament's original IMCO-LIBE report (20/04/2022) proposes language amending Annex III, adding 1.8.a, which would classify a large swath of content-generation systems as high-risk systems if they generate "text content that would falsely appear to a person to be human generated and authentic" or "audio or video content that appreciably resembles existing natural persons, in a manner that significantly distorts or fabricates the

original situation, meaning, content, or context and would falsely appear to a person to be authentic." Portions of this language overlap with Article 52's transparency obligations around disclosing that content has been artificially generated or manipulated, and we suggest aligning these requirements within Article 52 rather than adding a separate set of requirements under Annex III.

For more context, GPT-3 and our other general purpose AI systems such as DALL·E may generate outputs that could be mistaken for human text and image content. However, in line with the requirements outlined in Article 52, we require deployers building on our API to not mislead users that they are interacting with an AI system or AI generated content. We have developed mechanisms to allow us to verify the synthetic origin of images generated by DALL·E, and are constantly testing and iterating on restrictions within our Content Policy to address concerns around deepfakes and artificially generated content. For example, we currently prohibit the generation of images of specific individuals, but are exploring mitigations that we think would support benign use cases for such generations. We continue to evaluate methods to combat deepfakes and similar problems, and with current safeguards in place, we believe users will be aware that they are interacting with an AI system and that GPT-3 or DALL·E output does not mislead people.

Despite these efforts, the new language in Annex III 1.8.a could inadvertently require us to consider both GPT-3 and DALL·E to be inherently high-risk systems since they are theoretically capable of generating content within the scope of the clause. We suggest that instead of adding these additional clauses to Annex III, Article 52 can be relied on (or amended if deemed appropriate). This Article can sufficiently require and ensure that providers put into place reasonably appropriate mitigations around disinformation and deepfakes, such as watermarking content or maintaining the capability to confirm if a given piece of content was generated by their system.

III. Requiring New Conformity Assessments for Substantial Modifications

The AIA currently requires a new conformity assessment each time an AI system undergoes a "substantial modification", defined as a change that "affects the compliance of the AI system with the requirements set out in Title III Chapter 2 of this Regulation or results in a modification to the intended purpose for which the AI system has been assessed." We are concerned that this requirement may impact innovations that increase the safety of the AI system on the market, such as those achieved through our iterative deployment model. This model allows us to constantly reassess features and risk levels and make safety and security changes to our systems on a frequent, ongoing basis.

We propose that modifications made to increase the safety of an AI system on the market or to

mitigate risk should not be captured by “substantial modification”. For example, addressing concerns around hate speech may require monitoring the changing landscape of what constitutes hate speech (such as in relation to new social movements) and quickly updating systems accordingly. OpenAI’s iterative deployment allows our researchers and engineers to make improvements to our AI systems and tools to help ensure that they are continuously becoming safer, less biased, and more useful. This reduces the time between the discovery of important safety updates and the implementation and availability of such updates.

However, the current definition of “substantial modification” could be interpreted to require a new conformity assessment whenever changes such as these are made, as there is the possibility that it could affect the compliance of the AI system with broader Title III Chapter 2 requirements.[2] To avoid an undesirable outcome where improvements to the safety and well-functioning of an AI systems are unnecessarily delayed, we suggest excluding modifications made for safety or risk mitigation reasons that are not reasonably expected to have a negative impact on health, safety, or fundamental rights of any person; however, if the provider subsequently has reason to believe that such impacts have happened or may be likely, the modification should be rolled back and a new conformity assessment required before the modification is redeployed.

IV. Concerns With Scope of Certain High Risk Use Cases

Our final suggestions focus on specific categories of high risk use cases listed in Annex III. As mentioned earlier, OpenAI generally disallows most use cases deemed high risk by the AIA. However, there is some ambiguity where Annex III may capture certain low risk use cases. We believe it is critical that sectors fundamental to human growth and improvement, such as education and employment, are able to benefit broadly from AI advancements, particularly when the advancements do not pose a risk to a person’s fundamental rights.

As one example, Section 4.a in Annex III outlines “AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates.” While we agree that an AI system used to make or relied on as a primary input for direct employment decisions should be considered high risk, there are a number of use cases that could inadvertently be captured by the current language which would help benefit and modernize the sector without presenting risk to people. For example, one OpenAI customer uses GPT-3’s text generation capabilities to help people create and edit job descriptions for more effective recruiting. This could be considered as falling under the category of “advertising vacancies” in Section 4.a, but does not seem to be the primary

[2] We understand that since we operate primarily as a general purpose system provider, the conformity assessment may not be implicated at all. However, we or users may build applications that do fall under high-risk categories, in which case AI safety and risk mitigation efforts would be slowed by conformity assessment requirements for substantial modifications.

thrust of the categorization, because the AI system supports the human decision maker as an assistant and is not the primary author of the job description. The description is reviewed and ultimately finalized by a person, and the posting and availability of the description is not determined by the AI system.

Similarly, Section 3.b covers "AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions." As with the previous job description example, we believe that the use of our AI systems to, for example, help test writers develop and edit new test questions, would present large benefits in helping to modernize the educational sector without entailing risks as those intended to be addressed by Annex III. However, the specific language could be read to include the development of test questions to be part of “assessing students” or “assessing participants in tests.” We suggest that the use of AI systems to generate test questions for human curation, editing, and selection should not be considered high risk, as this would result in reduced ability for educators to benefit from AI advancements. To address these concerns, we propose that Sections 3 and 4 of Annex III be amended to clarify a focus on use cases that will have a material impact on a person’s employment or educational opportunities, and not potentially capture broad sector-wide use like test question generation or job description generation that present relatively low risk. An example of alternative language for potential consideration:

“3. Education and vocational training:

(a) AI systems intended to be used to make decisions regarding access to or assignment of natural persons to educational and vocational training institutions;

(b) AI systems intended to be used for evaluating performance in educational and vocational training institutions, including performance in tests commonly required for admission to educational institutions.

4. Employment, workers management and access to self-employment:

(a) AI systems intended to make decisions regarding the suitability of natural persons for employment, including determining access to job vacancies, screening or filtering applications, and evaluating candidate performance;

(b) AI systems intended to make decisions on promotion and termination of work-related contractual relationships, and for monitoring and evaluating performance and behavior of persons in such relationships.”

Additionally, given the continued advancement of AI systems' capabilities, we expect that currently unknown high risk use cases will continue to emerge, making it important to ensure that the AIA remains agile in capturing ongoing developments. Quickly capturing new high risk AI systems and removing those which have proven themselves sufficiently low risk must be low friction. We agree with the submissions that advocate for a process that can ensure a speedy turnaround when it comes to adding new high risk AI systems to Annex III. Equally, we welcome the Czech Presidency of the Council's proposal (15/07/2022), which includes an amendment to Article 7.3 empowering the European Commission to delete AI systems from Annex III via delegated acts under specific circumstances.

Conclusion

We hope that these comments provide a helpful perspective of the capabilities and safety mechanisms of general purpose AI systems, and we appreciate the opportunity to share some of our core expertise and viewpoints on the AIA. We recognize and appreciate the enormity of the EU's work in understanding and encouraging development of critical AI technology while ensuring that the development and use of these systems respects fundamental human rights and values. We remain ready to assist and advise however needed.