2 August 2021

OpenAI
3180 18th St
San Francisco, CA 94110

Comments on European Union Proposal on AI Rules and Actions

This submission from OpenAI provides feedback on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts. We previously submitted feedback on the White Paper on Artificial Intelligence and feedback on the Ethics Guidelines for Trustworthy AI's Technical Robustness and Safety section; we will draw from our previous feedback throughout this response. We offer recommendations for consideration when taking action towards shaping trustworthy and responsible artificial intelligence (AI). We aim to help the EU technically inform regulations, so they are most applicable to existing and future AI systems. If a section is not highlighted, we do not have specific, actionable feedback.

About OpenAI

OpenAI is an AI research and deployment company based in San Francisco whose mission is to develop transformative artificial intelligence and ensure it benefits all of humanity. OpenAI's work is primarily built around three areas: technical capabilities research and development; AI safety research and development; and policy work, which supports informed AI policymaking.

Toward a Legal Framework for Trustworthy AI

We congratulate the European Commission on this Proposal and support regulation on AI that guides AI use towards beneficial applications and away from potential risks and harms. We also broadly agree with the risk-based approach taken in this Proposal. Parts of AI research, development, and deployment will require government oversight and regulation. That oversight should encourage innovation in systems that enhance human quality of life.

Encompassing General-Purpose Systems

General-purpose AI systems can have many use cases and applications. A single system can meet the Act's definition of "unacceptable", "high-risk, and "low-risk". We encourage the European Commission to further consider how such systems will be effectively evaluated by the proposed rules. One means is by requiring a full risk assessment of an AI system's possible uses instead of categorizing AI systems by a limited set of stated intended purposes. This should also consider the societal impact of each use. We strongly urge against banning a system altogether simply because one of its many use cases falls under an unacceptable category. Instead, effective safeguards such as transparency requirements should be imposed and enforced to steer these systems toward only the beneficial applications.

An example of a general purpose system is OpenAI's language model GPT-3: an AI system that is able to predict the next word in a string of text, given a set of words as input. OpenAI released

GPT-3 through an API, where customers and researchers can apply and study the system respectively. GPT-3 powers hundreds of beneficial applications across a range of industries and use cases, from storytelling to semantic search. However, GPT-3 can be misused; our research partners at Middlebury Institute's Center on Terrorism, Extremism, and Counterrorism found that GPT-3 can accurately emulate extremist content and could be a tool to radicalize individuals into violent far-right extremist ideologies. OpenAI takes strict safety measures against potential malicious uses of GPT-3, such as limiting access to the API and monitoring use. Our effective safeguards show that a general-purpose system capable of high-risk behavior can be steered toward constructive, low-risk applications.

Transparency Obligations for Certain Systems

We support the proposed requirement that natural persons be informed that they are interacting with an AI system. We further support the obligation that generated images and audio be labeled as AI-generated or manipulated. Among other benefits, labeling AI outputs can help prevent disinformation. OpenAI's API Terms of Use enforces this obligation already, in order to prevent user misinterpretation of an output. Our research has shown that advanced AI system outputs can easily pass as human-generated. We also found that technical detection methods alone cannot be a fully accurate means of identifying an AI output. Proactively labeling outputs can prevent harmful social impact.

Key Terms for Clarification

We found three key themes in the Proposal that would be most effective with further clarification: manipulation risks posed by AI systems; what constitutes "high quality" datasets; and processes to ensure fairness in AI. Firstly, guidelines, or checklists, that help developers determine whether a system is capable of manipulation can help drive AI development away from these use cases. Secondly, the quality of a dataset is normative and difficult to objectively evaluate. We encourage setting EU standards for what qualifies a "high quality" dataset, or sharing what are desirable characteristics (e.g. proven consent from subjects). Thirdly, determining a system's fairness is technically difficult. Technical evaluations are continually evolving. We strongly support public funding for research and evaluations to help measure fairness and mitigate harms. These evaluations should also update iteratively and importantly should empower voices that are traditionally underrepresented.

Iterative Updating

AI progress moves fast. Development of new systems and applications often can become public sooner than policy action is able. We encourage iterative processes, like regular reviews of existing potential future systems and applications, to update the risk-based framework. It is critical to keep abreast of advancements and use cases. The proposed European Artificial Intelligence Board is well-positioned to conduct these reviews and updates. The Board should be authorized to propose updates to Annexes including the list of prohibited and high-risk systems. Furthermore, more resources should be given to this iteration; implementation will require many experts across overlapping technical and nontechnical fields. Ultimately,

government agencies must build more technical and interdisciplinary capacity for AI oversight and monitoring.

Conclusion
This Proposal is a timely and necessary step toward ensuring AI systems fulfill trustworthy requirements and protect fundamental human rights. We look forward to further supporting the European Commission with our technical expertise.

Submitted by:
Irene Solaiman, Public Policy Manager