OpenAI

1960 Bryant Street
San Francisco, CA 94110

Friday, August 11, 2023

Comment on DISR Paper on Safe and Responsible AI

OpenAI appreciates the opportunity to provide comments on the "Safe and Responsible AI in Australia" discussion paper shared by the Department of Industry, Science, and Resources (DISR).

As the DISR recognizes in its report, AI presents significant economic and social opportunities for Australia, including through applications in healthcare, engineering, and legal services. Despite these potential benefits, AI also brings considerable challenges, including risks of misuse, inaccuracies, and potentially dangerous capabilities.

We welcome Australia's framing of this discussion in terms of AI safety, accountability, harmonization, and risk mitigation. We agree that "a coordinated and coherent response from the government to emerging issues," one based on transparency, evaluation, and accountability, would best serve the dual goals of innovation and safety. We believe coordination is critical not only amongst domestic stakeholders, but also internationally, especially where highly capable foundation models are concerned.

AI regulation takes time, and the need for guidance and accountability is urgent. In order to help bridge the gap until new laws and policies are implemented, we have joined other leading labs in making voluntary commitments on safety, security and trust. We are also coordinating with other key stakeholders to form a new industry body, the Frontier Model Forum, to promote the safe and responsible development of frontier AI systems.

Risk-based Approaches

We strongly support a risk-based approach to AI regulation. We believe that a mature AI accountability regime will include both horizontal and vertical elements. That is, we both expect there to be some elements that apply to certain AI systems across domains of application, as well as some elements that are tailored to particular domains.

In our view, AI developers like us must act responsibly and take a careful and safety-focused approach regardless of the particular domains in which such models may be used.

A wide range of existing laws already apply to AI – including to our products – and the legal landscape is quickly evolving, with legislative and policy initiatives unfolding around the world. At the same time, long-established bodies of law, regulation, and other expectations in areas like medicine, education, and employment are already being interpreted and adapted in ways that will shape the role AI plays in those domains. We see these sector-specific efforts, informed by deep domain expertise, as a critical part of the AI accountability landscape.

The global adoption of AI with wide-ranging capabilities has accelerated rapidly. We anticipate that such systems will continue to become both more capable and more ubiquitous. This growth in capability and usage stands to offer significant economic benefits and assist humanity in tackling our most important and intractable challenges. At the same time, intentional misuse or errors of such increasingly powerful AI systems could result in significant harmful impacts far beyond those of the systems in use today. Given the cross-border impact of these systems, an effective governance mechanism that addresses the risks posed by future powerful AI technology requires international coordination.

We strongly support efforts to harmonize the emergent accountability expectations for AI both within and across countries. Australia, as a leading global economy and democracy, has a valuable contribution to make to establishing a cohesive international framework for AI governance. Important contributions to these efforts are being made by the US and UK governments, the European Union, the OECD, the G7 (via the Hiroshima AI process), G20, and others. Coordinating to align domestic and international accountability mechanisms would get us closer to meeting the challenge of building responsible AI that benefits humanity.

Highly Capable Foundation Models and Addressing Potentially Dangerous Capabilities

Highly capable foundation models can provide society with great benefits but also have the potential to cause harm. As the capabilities of these models get more advanced, so do the scale and severity of the risks they may pose, particularly if under direction from a malicious actor or if the model is not properly aligned with human values.

Rigorously measuring advances in potentially dangerous capabilities is essential for effectively assessing and managing risk. At OpenAI we are exploring and building various evaluations for potentially dangerous capabilities that range from simple, scalable, and automated tools to bespoke, intensive evaluations performed by human experts. We believe

dangerous capability evaluations are an increasingly important building block for accountability and governance in frontier AI development.

We also support the development of registration and licensing requirements for future generations of the most highly capable foundation models. Such models may have sufficiently dangerous capabilities to pose significant risks to public safety; if they do, we believe they should be subject to commensurate accountability requirements.

There remain many open questions in the design of registration and licensing mechanisms for achieving accountability at the frontier of AI development. We look forward to collaborating with policymakers in addressing these questions.

Voluntary Commitments to Build Public Trust

We welcome and encourage regulation of AI. At the same time we also acknowledge that in order to develop an effective regulatory framework, it is first essential to build a knowledge base about concrete governance practices in areas such as pre-deployment testing, content provenance, and trust and safety. This is why we support and are taking voluntary steps to build public trust, and to ensure that we are developing responsible AI that benefits humanity.

OpenAI has taken two important steps, along with other leading companies in the U.S., to demonstrate our commitment to developing responsible AI, and to build and inform the capacity for effective government regulation in the future.

Voluntary Commitments

In late July, OpenAI and other leading AI labs made voluntary commitments to reinforce the safety, security and trustworthiness of AI technology and our services. This process, coordinated by the White House, is an important step in advancing meaningful and effective AI governance, both in the US and around the world.

As part of our mission to build safe and beneficial AGI, we will continue to pilot and refine concrete governance practices specifically tailored to highly capable foundation models like the ones that we produce. We also continue to invest in research in areas that can help inform regulation, such as techniques for assessing potentially dangerous capabilities in AI models.

Frontier Model Forum

We have come together with other labs to establish a new industry body called the Frontier Model Forum that will draw on the technical and operational expertise of its member companies to benefit the entire AI ecosystem. Its anticipated activities include advancing technical evaluations and benchmarks, and developing a public library of solutions to support industry best practices and standards.

While AI offers tremendous promise to benefit the world, appropriate guardrails are required to mitigate risks. Important contributions to these efforts have already been made by governments and international organizations. To build on these efforts, further work is needed on safety standards and evaluations to ensure frontier AI models are developed and deployed responsibly. The Forum will be one vehicle for cross-organizational discussions and actions on AI safety and responsibility.

Essential Elements of AI Accountability for Highly Capable Foundation Models

Transparency

We share the DISR's emphasis on the importance of transparency across the AI life cycle. At OpenAI, a key part of our approach to accountability is compiling and publishing information about new AI systems that we deploy. Our approach draws inspiration from previous research work on model cards and system cards. To date, OpenAI has published two system cards: the GPT-4 System Card and DALL-E 2 System Card.

We believe that in most cases, it is important for such disclosures to analyze and describe the impacts of a system – rather than focusing solely on the model itself – because a system's impacts depend in part on factors other than the model, including use case, context, and real world interactions. Likewise, an AI system's impacts depend on risk mitigations such as use policies, access controls, and monitoring for abuse. We believe it is reasonable for external stakeholders to expect information on these topics, and to have the opportunity to understand our approach.

Qualitative and Quantitative Model Evaluations

Red teaming is the process of qualitatively testing our models and systems in a variety of domains to create a more holistic view of the safety profile of our models. We conduct red-teaming internally as part of model development, as well as with people who operate independently of the team that builds the system being tested. In addition to probing our organization's capabilities and resilience to attacks, red teams also use stress testing and boundary testing methods, which focus on surfacing

edge cases and other potential failure modes with potential to cause harm.

Red teaming is complementary to automated, quantitative evaluations of model capabilities and risks that we also conduct. It can shed light on risks that are not yet quantifiable, or those for which more standardized evaluations have not yet been developed. These evaluations enable model comparisons, facilitate safety research methodologies, and provide crucial input for decision-making on model deployment. Existing evaluations cover topics such as erotic content, hateful content, self-harm-related content, and more, measuring the likelihood of our models generating such content.

Conclusion

We at OpenAI support multi-stakeholder approaches to building safe and responsible AI in Australia. We encourage a combination of approaches, emphasizing coordination with international governance frameworks and greater harmonization amongst domestic laws. We understand that both vertical and horizontal frameworks may be appropriate and that establishing an AI governance regime takes time. We therefore recommend and encourage the adoption of voluntary commitments and complementary initiatives such as the Foundation Model Forum, to make progress on accountability while government-led legislation and approaches are developed.

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