CUSPE Cambridge University Science and Policy Exchange

CJSP Cambridge Journal of Science & Policy

Open-source provisions for large models in the AI Act

COMMUNICATION | EDITORIAL | SPECIAL FEATURE | **PERSPECTIVE** | REPORT | REVIEW

Harry Law[1,2] and Sébastien Krier[1]
[1] Google DeepMind, [2] University of Cambridge
Edited by Lukas Gunschera

Image from Adobe Stock Images, selected by Lia Bote

ABSTRACT

On 21 April 2021, the European Commission tabled a proposal for the European Union's Artificial Intelligence Act to introduce a common regulatory and legal framework for the development and deployment of artificial intelligence (AI). Subsequent amendments have sought to include generalist or 'foundation' models, including in some commercial contexts those released on an open-source basis. Critics have argued that such a move would harm smaller AI developers,

who rely on open-source practices to develop new products and services. Others contend that targeted measures are necessary, given the risks of misusing open-source systems. This paper focuses on approaches to open-sourcing foundation models in the context of the Act, rather than questions relating to general open-source practices or publication norms for AI systems. It seeks to summarise the movements related to open-source models in the Artificial Intelligence Act and introduces possible avenues for compromise.

SCIENCE ⇒ POLICY

While the European Union's AI Act continues to evolve to reflect a rapidly shifting external environment, legislators are—at the time of writing—set to place significant liabilities on the providers of open-source foundation models. While we believe that the burden of liability of open-source providers should be minimised in the short term, we propose that the release of today's open-source models should include testing and safety evaluations for dangerous capabilities. In the long term, however, it is prudent for regulators to design provisions concerned with preventing the proliferation of future more powerful, dangerous models released on an open-sourced basis.

Key Words

Artificial Intelligence Act, AI Act, general purpose AI systems, foundation models, GPAIS, open-source, artificial intelligence, AI, machine learning

The views expressed here are the authors' personal views and should in no way be attributed to Google DeepMind.

The European Commission proposed the European Union's Artificial Intelligence Act on 21 April 2021 to introduce a common regulatory and legal framework for the development and deployment of artificial intelligence (AI). The Act, which is expected to become law by 2024, groups applications of AI into four categories: unacceptable risk, high risk, limited risk, and minimal or no risk. High-risk systems are those AI systems: (i) whose deployment poses a significant risk of harm to the health, safety or fundamental rights of persons and fall under certain critical use cases or areas, such as critical infrastructure, education and essential private or public services; or (ii) that are safety components of a product or are themselves products that fall within certain existing EU product safety laws and require third party conformity assessment under those existing laws. Such systems are faced with mandatory requirements surrounding their development and deployment, including stipulations for introducing a risk management system, a quality management system, record keeping, efforts to govern systems' use of data, the introduction of technical documentation, and a commitment to ensure oversight of the system by a human.

The European Parliament's proposed compromise amendments from May 2023 also impose some requirements on 'foundation models', 'generative AI systems', and 'open-source AI systems'. Such a response may seek to address upstream developer obligations, content-related issues such as privacy and copyright, and the allocation of obligations along the AI value chain (including apportioning liability between system providers and deployers). For the purposes of this paper, and in the interest of clarity and brevity, we will focus on 'foundation models' - defined as an "AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks."

Advances in AI research have continued since the Act was first proposed, typified by the rise in large generative models and the

development of more generalist agents. Interest in foundation models has sharply increased following the release of major language models and applications based on them such as ChatGPT (and more recently, GPT-4) from San Francisco-based developer OpenAI. The popularity of these highly generalisable systems has been buoyed by open-source versions of text-to-text models such as the Falcon LLM model released by the Abu Dhabi-based Technology Innovation Institute and text-to-image models, including Stable Diffusion built by London-based Stablity.AI [1]. Amidst the growing popularity of such systems, lawmakers considered in the second half of 2022 whether general-purpose AI systems or GPAIS—a loosely defined group of AI systems that have or enable a broader set of capabilities—ought to be included within the high-risk category [2]. This has proven controversial, as the changes targeted a specific type of model, contravening the Act's risk-based approach. As stated above, recent drafts also refer to 'foundation models' and other definitions. The original scope of the AI Act was intended only to cover AI systems deployed in high-risk settings, such as tools designed for cancer detection. But subsequent amendments regarding general-purpose or foundation models have targeted their *development*. Open-source systems were included at certain points and excluded during other periods, which underscores the complexity of the question and the challenge it poses for policy development.

As of June 2023, certain open-source systems are currently in-scope: "A provider of a foundation model shall, prior to making it available on the market or putting it into service, ensure that it is compliant with the requirements set out in this Article, regardless of whether it is provided as a standalone model or embedded in an AI system or a product, or provided under free and open-source licences, as a service, as well as other distribution channels." This provision will apply to models made available in the course of commercial activity, or supplied for first use to a deployer or put into service for their own use. Recital 12b explains that a commercial activity can be characterised as "charging a price, with the exception of transactions between micro enterprises, for a free and open-source AI component but also by charging a price for technical support services, by providing a software platform through which the provider monetises other services, or by the use of personal data for reasons other than exclusively for improving the security, compatibility or interoperability of the software". The result of this settlement is that the AI Act is likely to place significant liabilities on some providers of open-source foundation models, in some specific circumstances.

Advocates of open-source approaches are concerned that possible attempts to regulate foundation models in this way would effectively ban the practice. Chief amongst the critics is Alex Engler, who wrote in August 2022 that 'while intended to enable the safer use of these tools, the proposal would create legal liability for open-source GPAI models, undermining their development [3].' This could have the unintended effect of, Engler notes, concentrating power in the hands of technology companies and preventing research that is critical to public understanding of AI. Hugh Zhang, for example, argues that public knowledge about model capabilities will help reduce the spread of misinformation [4]. Others are pushing for provisions that would remove such systems from the scope or exclude open-source foundation models from the Act entirely. A coalition of technologists and think tanks, however, has argued that for the AI Act to provide meaningful protections against high risk uses of AI systems, foundation models must be included within its scope [5]. Some commentators have suggested that this should include open-source systems too. Such perspectives hold that as models become more capable and powerful, a reckless developer may release an unsafe model—for example, with spyware

generation capabilities—leading to increases in risk with model proliferation and no associated liability or accountability [6]. While such models do not yet exist, there are legitimate reasons to believe that future models may present novel and dangerous capabilities that materially affect the scope and breadth of risk we are concerned with. As a result, it is prudent for regulators to design provisions concerned with preventing the proliferation of future more powerful, dangerous models released on an open-source basis.

One possible way forward, as described by Open Future's Paul Keller, is the introduction of transparency requirements for the systems. These requirements might include standardised information in the form of model cards [7] and system cards [8], access to the data used to train and fine-tune the model to enable auditing and, where possible, an explanation of the model's behaviours [9]. In fact Recital 12c currently specifies that open-source developers "should however be encouraged to implement widely adopted documentation practices, such as model and data cards, as a way to accelerate information sharing along the AI value chain, allowing the promotion of trustworthy AI systems in the EU." However, some transparency requirements proposed in the AI Act are either technically impossible to comply with, or very difficult to implement (e.g. "sufficiently detailed" summary of copyrighted training data). Additionally, even the most stringent transparency requirements may not address the risk of harmful models diffusing. As researchers at the Center for a New American Security have noted, the risk of models being used for malicious purposes by authoritarian regimes requires firms to develop appropriate security measures and refrain from open-sourcing the technical specificities of cutting-edge models [10]. So while transparency provisions may in theory be helpful in some respects, they do not practically address a potential future need to limit the spread of models that present 'dual-use' capabilities. Indeed, some researchers believe more capable models may in the future enable risks such as cyber-offense or extreme manipulation capabilities [11].

Open-source software and code remain important components for the maturation of the AI industry, with the sharing of notebook environments, libraries, datasets, and tools critical to the growth of AI startups and the acceleration of research. Open-source software also allows AI to be used and tested by millions of people around the world. Google DeepMind's AlphaFold, for example, is now fundamental to the work of many scientists attempting to cure malaria, understand cancer or limit antibiotic resistance. To prevent the formation of undue and unnecessary barriers to entry, any restriction should therefore be precisely targeted: for example, by looking at a certain threshold of capabilities, parameter size, or training time. How exactly this threshold can be precisely defined remains a contentious question. Although the vast majority of models and tools are unlikely to cause significant harm, we can reasonably expect new threat vectors to emerge as they grow in capability. Three years ago, models struggled with differentiating between cats and dogs. Today, large language models could help develop sophisticated malware (though not create it from scratch). As the UK's National Cyber Security Centre notes, while doing so would still require expert knowledge to validate and deploy, this hurdle will be gradually lowered as generative models improve [12]. White House officials have also recently noted that using open-source systems voluntarily maintained by the community may pose national security concerns, as the vulnerability found in Log4J (an open-source library commonly used by apps and services across the internet) demonstrated. Similar dynamics may well surface as companies and governments rely on open-source models [13].

Assuming the sophistication of models continues to develop at current rates, it may be necessary to review the diffusion and misuse risks associated with foundation

models. The full set of AI Act obligations may be disproportionate. However, two related requirements could be warranted: testing and safety evaluations for dangerous capabilities and sufficient access to verify those claims. More work must be done first to develop useful evaluations and testing tools, without which the definition of a powerful foundation model will likely be arbitrary. As a guide to the level of risk posed by an AI system, a designation as a 'general purpose model' or 'foundation model' is likely of limited use in isolation, without criteria to assess an AI system's relative generality, applicability, and viable uses. As Isabella Duan explains, safety-relevant benchmarks can support the empirical, quantitative evaluation of AI systems and promote healthy competition over safety-relevant properties [14]. Initiatives like Stanford's HELM framework to evaluate the performance of language models [15], ARC Evals that seek to understand the capabilities and safeness of large models [16] and the OpenAI Evals, which includes an open-source registry of model evaluations [17], are steps in the right direction.

This paper has described the core components of the AI Act with reference to provisions focused on managing the release of open-source models. It has discussed recent efforts to reach a settlement between the regulation of foundation models and general purpose models. Our analysis has connected open-source foundation models with provisions in the AI Act that we argue could place significant liabilities on providers. We have suggested that doing so would discourage the release of open-source foundation models, which may have a knock-on effect on the growth of the AI industry and the ability of users to understand the capabilities and limitations of models. Our account weighs this outcome against the ability of open-source developers to release powerful models whose deployment may cause significant harm. We argue that, while the release of today's open-source models should be supported, regulators ought to design provisions concerned with preventing the proliferation of powerful, dangerous models released on an open-source basis in the future.

In practice, we propose that the release of today's open-source foundation models should include testing and safety evaluations for dangerous capabilities to ensure policymakers and users of such models are aware of the risks associated with their release. The development of new benchmarks, metronomy and evaluations will enable a more precise delineation and targeting of powerful foundation models. At the point at which models start demonstrating more concerning capabilities, it may be the case that they should only be partly released, or in some cases not at all. Shevlane and Dafoe have noted that AI researchers can choose the components or artefacts they wish to share or restrict: basic results and insights, detailed results, code, trained networks, and easy-to-use tools [18]. Any future restrictions should therefore be targeted and address the root of the problem by, for example, permitting the sharing of specific components but not source code or weights. Shevlane suggests a structured access mechanism, for example, that seeks to introduce controlled arms-length interactions with AI systems [19]. A careful balance must be stuck, as an overly broad scope and set of requirements may inadvertently cement the market power of incumbents. Access mechanisms (e.g. through APIs) should allow researchers to verify claims by developers, whilst ensuring that a narrow set of highly capable foundation models do not proliferate before they are deemed safe. Ultimately, we contend that for open-source models the Act's broad scope risks encompassing too much. Instead, we suggest the development of better tools, evaluations, definitions, and thresholds to limit risks and negative externalities, and propose only restricting the release of open-source foundation models that are demonstrably more complex, capable, and dangerous.

References

[1] Alek Tarkowski, Open Future blog, 'Notes on BLOOM, RAIL and openness of AI', last accessed 14 January 2022, https://openfuture.pubpub.org/pub/notes-on-open-ai/release/1.

[2] Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. "On the Opportunities and Risks of Foundation Models." arXiv.org, July 12, 2022. https://arxiv.org/abs/2108.07258.

[3] Alex Engler, Brookings Institution blog, 'The EU's attempt to regulate open-source AI is counterproductive', last accessed 15 January 2022, https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/

[4] Zhang, Hugh. "Dear Openai: Please Open-source Your Language Model." The Gradient. The Gradient, December 30, 2021. https://thegradient.pub/openai-please-open-source-your-language-model/.

[5] Anon., Future of Life Institute, 'Call for better protections of people affected at the source of the AI value chain', last accessed January 15 2022, https://futureoflife.org/wp-content/uploads/2022/10/Civil-society-letter-GPAIS-October-2022.pdf

[6] Langenkamp, Max. "How Open-source Machine Learning Software Shapes Ai." maxlangenkamp.me. Accessed March 16, 2023. https://maxlangenkamp.me/posts/mloss_essay/.

[7] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru: "Model Cards for Model Reporting", 2018, FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA, USA; arXiv:1810.03993. DOI: 10.1145/3287560.3287596.

[8] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru: "Model Cards for Model Reporting", 2018, FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA, USA; arXiv:1810.03993. DOI: 10.1145/3287560.3287596.

[9] Paul Keller, Open Future blog, 'How will the AI Act Deal With Open-source AI Systems', accessed 14 January, https://openfuture.eu/blog/how-will-the-ai-act-deal-with-open-source-ai-systems.

[10] Bill Drexel and Caleb Withers, Opinion Contributors. "Generative AI Could Be an Authoritarian Breakthrough in Brainwashing." The Hill. The Hill, February 27, 2023. https://thehill.com/opinion/technology/3871841-generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing/.

[11] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, Allan Dafoe: "Model evaluation for extreme risks", 2023; arXiv:2305.15324.

[12] "ChatGPT and Large Language Models: What's the Risk?" NCSC. Accessed March 16, 2023. https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.

[13] "White House Enlists Software Industry to Improve Open-Source Security." Yahoo! Finance. Yahoo! Accessed March 16, 2023.

https://finance.yahoo.com/news/white-house-enlists-software-industry-213026544.html.
[14] Duan, Isabella. "Race to the Top: Rethink Benchmark-Making for Safe AI Development." Isabella Duan, December 3, 2022. https://isaduan.github.io/isabelladuan.github.io/posts/first/.
[15] Holistic evaluation of Language Models (Helm). Accessed March 16, 2023. https://crfm.stanford.edu/helm/latest/.
[16] ARC Evals. Accessed March 16, 2023. https://evals.alignment.org/.
[17] Openai. "Openai/Evals: Evals Is a Framework for Evaluating Openai Models and an Open-Source Registry of Benchmarks." GitHub. Accessed March 16, 2023. https://github.com/openai/evals.
[18] Shevlane, Toby, and Allan Dafoe. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" arXiv.org, January 9, 2020. https://arxiv.org/abs/2001.00463.
[19] Shevlane, Toby. "Structured Access: An Emerging Paradigm for Safe AI Deployment." arXiv.org, April 11, 2022. https://arxiv.org/abs/2201.05159.

About the Authors

Harry Law

Harry is an ethics and policy researcher at Google DeepMind and a History and Philosophy of Science PhD candidate at the University of Cambridge. His academic research examines the social, material, and political circumstances surrounding the development of artificial neural network technology in the United States. He is a postgraduate fellow at the Leverhulme Centre for the Future of Intelligence.

Corresponding address:
harrylaw@deepmind.com

Sébastien Krier

Séb is an International Policy Manager at DeepMind, where he explores governance mechanisms to better control frontier AI models. Prior to this role, he was a Senior Technology Policy Researcher at Stanford University's Cyber Policy Centre, focusing on the influence of emerging technologies on democracy, authoritarianism, and global powers.

Before joining Stanford, Séb advised various governments and organizations as a strategic advisor on AI policy and governance. His work in AI policy began at the UK Government's Office for Artificial Intelligence in 2018, where he contributed to shaping national the UK's approach to AI governance and regulation.

Previously, Séb practiced law at international law firms Bryan Cave Leighton Paisner and Freshfields Bruckhaus Deringer. He specialized in international disputes and human rights, managing complex cases related to international arbitration, investor-state disputes, and public international law.

Conflict of interest: Both authors are currently employed by Google DeepMind.