bruegel

ANALYSIS

07 March 2024

The European Union AI Act: premature or precocious regulation?

As it stands, it is unknown whether the Act will stimulate responsible AI use or smother innovation.

Bertin Martens

Governments around the world are creating regulation to come to grips with the perceived risks of Artificial Intelligence (AI). The United States issued an AI Executive Order [1] while the UK government released a non-binding Declaration of Principles [2]. China imposed a light-touch business-friendly AI regulation, primarily meant as a signal to accelerate technological progress (Zhang, 2024). The European Union's Artificial Intelligence Act was proposed by the European Commission in April 2021 and the agreed final version is set for formal approval in the European Parliament and Council in April 2024.

What does the EU AI Act aim to do?

The Act is essentially a product safety regulation designed to reduce risks for humans from the use of AI systems. Product safety regulation works for single purpose products; the risks from application for that purpose can be assessed. Many older-generation AI systems are trained for a single application. The problem comes with the latest general purpose Large Language Models and Generative AI systems like OpenAI's ChatGPT, Meta's Llama or Google's Gemini, which are models that can be molded for an almost infinite range of purposes. It becomes difficult to assess all risks and to design regulations for all possible uses. The AI Act tries to work around this with a general obligation to avoid harm to fundamental rights for humans. According to one of the co-architects of the Act in the European Parliament, this regulatory mix of product safety and fundamental rights criteria is not adapted to AI models [3].

The AI Act classifies AI systems used in the EU, irrespective of where they are developed, according to level of risk. Most AI applications are considered minimal risk and not regulated. Limited risk systems are subject to transparency and user awareness obligations only, like chatbots and the watermarking of AI media output. Meanwhile, systems that are deemed to pose unacceptable risks are prohibited. These systems include remote biometric identification and categorisation, facial recognition databases and social scoring – with exceptions for medical and security reasons, which are subject to judicial authorisation and the respect of fundamental rights. The bulk of the AI Act focuses on regulation of high-risk AI systems, in between limited and unacceptable risk. These are single- or limited-purpose AI systems that interact with humans in education, employment, public services etc. The Act contains a complex set of rules and requirements to assess whether and under what conditions high-risk systems can be used.

Besides high-risk AI systems, there are General Purpose AI (GPAI) models. This refers to Large Language and Generative AI foundation Models[4]. These are considered general purpose because they can be applied to a wide range of tasks. GPAI providers must present technical documentation and instructions for use, unless they are open license models that can be adapted by users for their own purposes. Data used for training must be summarily documented and must comply with the EU Copyright Directive. In the law, GPAI models become systemic risk models when the computing power used for their training exceeds 10 flops (floating point computer operations). Providers of systemic risk GPAI models must conduct model evaluations and adversarial testing, provide metrics used to avoid harmful applications, report incidents and ensure cybersecurity protection. Currently available models do not reach that threshold[5]. But next-generation AI models, which could possibly be released in 2024, are likely to exceed the threshold. Eventually, it may capture all new large AI models.

Fundamental human rights safeguards and risks in AI models

The AI community has put a lot of effort into human-centric AI and the 'alignment' of AI model responses with human values, avoiding discrimination and harmful responses. That chimes with the contemporary diversity, equity and inclusion debate that targets racial, gender, sexual and religious discrimination. Some companies go to great lengths in this respect. Google trained its Gemini AI model to prioritise racial diversity over historical correctness[6]. But there are many other discrimination criteria that are frequently used. For example, price and income discrimination may be either welfare enhancing or exploitative when used for better targeting of economic services and subsidies. A rule that allows or bans it will for sure make a mistake in one direction or the other. This raises questions: whose values, harms and benefits should we align with? AI is already being used in defense and warfare: is that a human-centric application?

GPAI developers try to ensure respect for human values by building 'guardrails' into models. However, there are many ways to circumvent these guardrails [7] , through poisoning attacks that manipulate training data to falsify outputs, introducing malicious code via pre-packaged prompts, sponge attacks that destabilise computing power in the AI system, inference attacks that reveal hidden data that it is not supposed to be disclosed or deception attacks by means of visual illusions that are invisible to the human eye. Some developers try to build 'constitutional' guardrails for models to self-check whether their responses comply with obligations [8] .

Open GPAI models are more prone to loopholes to circumvent guardrails. But openness may spur innovative applications and new revenue-generating business models that are especially important for smaller AI firms that do not have a well-established business setting where they can put their models to work. The AI Act leaves developers of open models off the hook, unless they represent systemic risks, by exonerating them from testing obligations and passing on that responsibility to re-developers and deployers who can modify the behaviour of these models. It is more difficult to track regulatory compliance when many layers of complementary application providers interact.

Smaller models also escape stringent AI Act obligations because they do not meet the computing power threshold. Besides lowering training costs, this exemption also reduces regulatory compliance costs. While they can be just as versatile as larger models in the range of applications, they usually give less accurate replies unless they receive additional guidance from prompts and user datasets. Smaller models therefore do not necessarily imply lower risks.

The AI Act obliges large model developers to explain to downstream deployers and service providers how a given model interacts or how it can be used to interact with hardware or software that is not part of that model. This is a very generic provision that will require further clarification through implementing acts. It raises intriguing questions about vertical and horizontal integration between GPAI models and complementary services and the responsibilities of these parties in the context of the AI Act. Who is the deployer when an online travel services platform pulls in an AI application to improve consumer services: the application provider or the platform? There may be several layers of deployers – an issue currently not covered by the AI Act but subject to interpretation in implementing guidelines.

The incompleteness of the AI Act fails to provide legal certainty to AI developers and deployers. Moreover, it generates high compliance costs, especially for SMEs and start-ups that might find the EU regulatory environment too costly and risky [9] . However, the Act sets the scene for further regulatory work for the European Commission and its newly created AI Office. The Office will register and verify notifications sent by AI developers. However, the Office will contend with limited resources and will take a while to get up to speed. It will have to produce more than a dozen detailed implementation acts and guidelines, including delegated acts on the definition of AI systems, clarifying further the criteria for high-risk AI systems, adapting thresholds for general purpose AI models with systemic risk, specifying technical documentation requirements for general purpose AI and conformity assessments as well as issuing a code of practice for providers of general-purpose AI models. Moreover, it can clarify prohibited AI practices, requirements for high-risk systems and transparency obligations *"when deemed necessary"*. This may expand or tighten the regulatory space within which AI model developers can operate in the EU.

The AI Act and competition

There are by now several dozen large AI foundation models and many smaller models. Numbers are growing exponentially. There is vigorous competition between AI developers but no sign of emerging monopolistic gatekeepers yet, except perhaps at the level of AI computing infrastructure where big tech firms clearly dominate.

The EU is currently not home to very large AI models. Regulators may count on the ‘Brussels effect’ of the AI Act: if other countries adopt similar regulations, it will level the competitive playing field and weaken incentives for developers to circumvent regulatory compliance costs and move elsewhere. Moreover, a Brussels seal of confidence may make an AI model more attractive and competitive. However, stringent and costly regulatory conditions for large models may further entrench small AI developers in a niche market for smaller models that remain below the regulatory radar for systemic risks but also far away from the technology frontier.

The competition policy implications of AI technologies are not clear at this stage. The development and training of frontier models costs hundreds of millions of dollars and is often beyond the reach of AI start-ups. Established big tech firms including Google, Meta, Microsoft and Amazon have leveraged their extensive cloud computing infrastructure to develop their own large AI models.   Start-ups, often created by former big tech employees, are more innovative and closer to the AI technology frontier. However, they require close collaboration agreements with large tech firms that make expensive computing infrastructure and data available in return for access to the model. The recent agreement between the French AI startup Mistral and Microsoft illustrates this [10] . These agreements stop short of mergers. Competition authorities have started looking into the nature of collaboration between OpenAI and Microsoft [11] . To what extent smaller AI models, that are sufficiently performant for a wide range of tasks, can compete with larger firms and models remains an open question.

The AI Act has a narrow focus on regulation of self-standing large AI models. But the rapid emergence of decentralised AI ecosystems, whereby AI models are increasingly interacting with complementary and competing platforms and software through apps and plugins means systemic risks cannot be assessed by focusing on a single model; one needs to look at the whole system. Risks are shifted between developers, deployers and users. How much vertical and horizontal complementarity and integration between models and other system components can be allowed before it distorts markets?

There is a booming ecosystem of purpose-built ChatGPT applications [12] that enhance model performance with extensive sets of natural language prompts and propriety datasets that guide it towards answers in specialised domains. GPAI models are becoming operating systems on top of which deployers and end users can build their own apps for specific applications. Plugins enable it to connect to existing platforms and software, like to explore travel, e-commerce and other services, or perform specialised calculations and services. AI app stores may eventually overtake smartphone app stores with all-purpose services apps.

AI and copyright

Training models require massive data inputs, including text harvested from webpages, scanned documents and books, images collected from the web, video from film archives, sound from music collections. Much of this is subject to copyright. The EU Copyright Directive[13] includes an exception to copyright protection for text and data mining purposes, provided the user has legal access to the inputs (ie, no hacking of paywalls for example). Copyright holders can opt-out of the exception and charge fees. Several news publishers have reached licensing agreements with big tech AI developers who can afford the fees. AI start-ups are waiting for the outcome of several pending court cases that should clarify the interpretation of copyright law for AI applications, including the application of the 'transformative use' copyright doctrine in the US. Licensed datasets may be of higher quality and reduce training costs. But they may also reduce the set of available training data and result in biased training. Moreover, granting copyright on training inputs gives the private interests of copyright holders leverage over the wider social welfare implications of AI models that are rapidly becoming a general-purpose technology that is used across all sectors in the economy, far beyond the creative media industries that have a private interest in copyright.

Creative artists who use AI start to claim copyright on outputs. The AI Act states that AI audio-visual and text outputs should have machine readable watermarks to distinguish them from human outputs and deepfakes. Watermarking technology is still in the early stages and easily subject to circumvention[14]. The watermarking obligation does not apply when AI only assists humans. In most countries, only human outputs can claim copyright, not machine outputs. How much human contribution is required in a hybrid output to claim copyright? A single line of human-written 'prompts' may not be enough, but a Chinese court recently granted copyright to a developer of a complex set of prompts (Wang and Zhang, 2024). These issues show the EU Copyright Directive may need some re-thinking[15].

The EU AI Act as it stands is just the start of a long regulatory process. It delegates responsibility to the Commission and its newly created AI Office to draft implementation acts and guidelines to address these challenges. These will drive enforcement of the Act and determine to what extent it will be a precocious instrument to stimulate trustworthy AI innovation or a premature innovation-smothering regulation.

References

Wang, Y. and J. Zhang (2024) 'Beijing Internet Court Grants Copyright to AI-Generated Image for the First Time', *Kluwer Copyright Blog,* available at https://copyrightblog.kluweriplaw.com/2024/02/02/beijing-internet-court-grants-copyright-to-ai-generated-image-for-the-first-time/

Zhang, A. (2024) 'The Promise and Perils of China's Regulation of AI', *University of Hong Kong Faculty of Law Research Paper* 2024/02, available at https://ssrn.com/abstract=4708676

Endnotes

1. See https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

2. See https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

3. Kai Zenner, 'Some personal reflections on the EU AI Act: a bittersweet ending', 16 Feb 2024, available at https://www.linkedin.com/pulse/some-personal-reflections-eu-ai-act-bittersweet-ending-kai-zenner-avgee">

4. UK Competition and Markets Authority (2023) AI Foundation Models: an initial review, available at https://www.gov.uk/cma-cases/ai-foundation-models-initial-review

5. OpenAI's ChatGPT-4 is estimated at 2x1025 flops, Meta's Llama 2 at 'only' 8x1023 flops.

6. Madhumita Murgia, 'Google pauses AI image generation of people after diversity backlash', The Financial Times, 22 Feb 2024, https://www.ft.com/content/979fe974-2902-4d78-8243-a0cff68e630a

7. See https://www.vischer.com/en/knowledge/blog/part-6-the-flip-side-of-the-coin-where-we-need-to-protect-ai-from-attackers/

8. See https://huggingface.co/blog/constitutional_ai

9. [9] Kai Zenner, op.cit.

10. See https://azure.microsoft.com/en-us/blog/microsoft-and-mistral-ai-announce-new-partnership-to-accelerate-ai-innovation-and-introduce-mistral-large-first-on-azure/

11. See press release 'Commission launches calls for contributions on competition in virtual worlds and generative AI', available at https://ec.europa.eu/commission/presscorner/detail/en/ip_24_85

12. See https://openai.com/blog/introducing-the-gpt-store

13. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market.

[14] Kate Knibs, 'Researchers Tested AI Watermarks—and Broke All of Them', Wired, 3 October 2023, https://www.wired.com/story/artificial-intelligence-watermarking-issues/

[15] Bruegel hosted a debate on this. Streaming available at https://www.bruegel.org/event/eu-copyright-directive-still-fit-purpose-age-generative-ai

Bruegel
Rue de la Charité 33,
B-1210 Brussels
(+32) 2 227 4210
info@bruegel.org

Bruegel.org