CSET

The EU AI Act: A Primer

Mia Hoffmann

September 26, 2023

The EU AI Act is nearing formal adoption and implementation. Read this blog post, with updated analysis following the December 2023 political agreement, by CSET's resident EU expert and Research Fellow, Mia Hoffmann. Learn what we know about the Act and what it means for AI regulation in the EU (and the world).

[Note from Author: Updated January 4, 2024]: On December 8, 2023, the European Parliament and the Council of the EU, as part of the trilogue process explained further below, reached a political agreement on the AI Act <https://www.euractiv.com/section/artificial-intelligence/news/european-union-squares-the-circle-on-the-worlds-first-ai-rulebook/>. The provisional agreement concludes many years of

preparation, development and negotiation on the most comprehensive regulatory framework on AI to date, and presents a major step towards final adoption of the regulation.

Until the compromise text is published, the details of the final governance regime are unclear. Known key differences to the provisions and requirements described in the original September post below relate to the governance of general-purpose AI (GPAI) models. Similar to the use-case approach, GPAI systems will also be grouped into risk tiers, with those believed to pose systemic risks subject to more stringent obligations like model evaluations, risk management practices and incident reporting. All GPAI must comply with transparency requirements concerning their development and testing, energy consumption and training data.

Technical meetings are underway to consolidate the text which is expected to be formally adopted by spring. This will launch a longer rule-making and implementation process, as described below, with most provisions not coming into effect until at least 2025. Our original analysis below, from September 2023, remains a useful summary of the EU AI Act. Once the full text is available, we will likely further refine this primer.

Original Post

[Posted September 26, 2023] The *Proposal for a Regulation laying down harmonised rules for artificial intelligence*, better known as the EU *AI Act* will be finalized by the end of the year. Pending final EU procedures (the *trilogue)*, the act will likely be adopted in early 2024 before June 2024 European

Parliament elections. Its enactment will be followed by a transition period of at least 18 months before the regulation becomes fully enforced.

This blog offers a high-level introduction to the AI Act for those interested in AI regulation, and a refresher for readers who lost track of the deliberations over the last two years. It provides an overview of the core concepts and approaches of the AI Act, focusing on the commonalities and small differences between the proposals of the Council of the EU <https://data.consilium.europa.eu/doc/document/st-14954-2022-init/en/pdf>, the European Parliament, <https://www.europarl.europa.eu/doceo/document/ta-9-2023-0236_en.html> and the European Commission that will be hammered out during the trilogue process over the coming months. However, there remain substantive disagreements in the proposals, especially when it comes to technical and detailed provisions like definitions and enforcement. Dissecting those warrants a blog post on its own and is not the focus of this article.

What is the EU AI Act?

The AI Act is a legal framework governing the sale and use of artificial intelligence in the EU. Its official purpose is to ensure the proper functioning of the EU single market by setting consistent standards for AI systems across EU member states. In practice, it is the first comprehensive regulation addressing the risks of artificial intelligence through a set of obligations and requirements that intend to safeguard the health, safety and fundamental rights of EU citizens and beyond, and is expected to have an outsized impact on AI governance worldwide.

The AI Act is part of a wider emerging digital rulebook in the EU that regulates different aspects of the digital economy like the General Data Protection Regulation <https://commission.europa.eu/law/law-topic/data-protection/eu-data-protection-rules_en#library>, the Digital Services Act <https://digital-

strategy.ec.europa.eu/en/policies/digital-services-act-package>, and the Digital Markets Act <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>. As such, the AI Act does not address data protection, online platforms or content moderation. While the interplay between the AI Act and existing EU legislation poses its own challenges, building on existing laws enables the EU to avoid a "one law fixes all" approach to this emerging technology.

The AI Act covers AI systems that are "placed on the market, put into service or used in the EU." This means that in addition to developers and deployers in the EU, it also applies to global vendors selling or otherwise making their system or its output available to users in the EU.

There are three exceptions:

- AI systems exclusively developed or used for military purposes, and possibly defense and national security purposes more broadly, pending negotiations;
- AI developed and used for scientific research; and,
- Free and open source AI systems and components (a term not yet clearly defined), with the exception of foundation models which are discussed below.

The Risk-Based Approach

At the heart of the proposal stands its risk categorization system, whereby AI systems are regulated based on the level of risk they pose to the health, safety and fundamental rights of a person. There are four categories of risk: unacceptable, high, limited and minimal/none. The greatest oversight and regulation envisioned by the AI Act focuses on the unacceptable and high risk

categories, so that is the focus of much of the discussion below. Exactly where different types of AI systems fall remains to be determined and is expected to be fiercely debated during the trilogue. In practice, an AI system might also fall into several categories.

Unacceptable Risk Systems will be Prohibited

AI systems belonging to the unacceptable risk category are prohibited outright. Based on consensus between the three proposals, unacceptable risk systems include those that have a significant potential for manipulation either through subconscious messaging and stimuli, or by exploiting vulnerabilities like socioeconomic status, disability, or age. AI systems for social scoring, a term that describes the evaluation and treatment of people based on their social behavior, are also banned. The European Parliament further intends to prohibit real-time remote biometric identification in public spaces, like live facial recognition systems, alongside other biometrics and law enforcement use cases.

High Risks Systems will be Carefully Regulated

High-risk AI systems fall into one of two categories:

1. System is a safety component or a product subject to existing safety standards and assessments, such as toys or medical devices; or,

2. System is used for a specific sensitive purpose. The exact list of these use cases is subject to change during the negotiations, but are understood to fall within the following eight high-level areas:
    - Biometrics
    - Critical infrastructure
    - Education and vocational training
    - Employment, workers management and access to self-employment
    - Access to essential services
    - Law enforcement
    - Migration,asylum and border control management
    - Administration of justice and democratic processes

While the Council's proposal introduces additional exemptions for law-enforcement use, the European Parliament instead proposes a broader set of high-risk use cases. It adds, for example, content recommender systems of large online platforms, like social media algorithms, and AI systems used for the detection and identification of migrants. At the same time, the proposal adds a qualifier: In addition to its use in a critical context, an AI system must pose a significant risk of harm in order to fall under the high-risk category. Further guidance on the circumstances under which a system does or does not meet this threshold would be issued after the regulation's adoption.

Requirements for High Risk AI Systems

Under the proposals, developers of high-risk AI systems must meet various requirements demonstrating that their technology and its use does not pose a significant threat to health, safety and fundamental rights. These include a

comprehensive set of risk management, data governance, monitoring and record-keeping practices, detailed documentation alongside transparency and human oversight obligations, and standards for accuracy, robustness and cybersecurity. High-risk AI systems must also be registered in an EU-wide public database.

Developers determine their AI system's risk category themselves. Similarly, and with a few exceptions, developers may self-assess and self-certify the conformity of their AI systems and governance practices with the requirements described in the AI Act. They can do so either by adopting forthcoming standards <https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/> or by justifying the equivalency of other technical solutions. Miscategorizing an AI system and/or failing to comply with the respective provisions is subject to a fine of at least 20 million Euro or 4% of global turnover, whichever is higher (numbers subject to change during the trilogue).

Following developers' certification of conformity, deployers are required to comply with monitoring and record-keeping, as well as human oversight and transparency obligations once they put a high-risk AI system to use. The Parliament is also pushing for deployers to conduct a fundamental rights impact assessment, recognizing that risks are context-dependent. A signature verification system for renewing drivers licenses and one used to verify mail-in ballots in an election have vastly different implications. Even a well-working and technically safe AI system may not be appropriate under some circumstances.

Incident Reporting

Even state-of-the-art testing protocols and well-implemented risk management strategies will not entirely prevent harm from occurring once AI systems are deployed. Tracking AI incidents and learning from them to improve AI design, development and deployment is a crucial part of any safety strategy, as we argue in a recent paper <https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>.

The AI Act requires developers of high-risk AI to set up a reporting system for serious incidents as part of wider post-market monitoring. A serious incident is defined as an incident or a malfunction that led to, might have led or might lead to serious damage to a person's health or their death, serious damage to property or the environment, the disruption of critical infrastructure or the violation of fundamental rights under EU law. Developers and, in some cases, deployers must notify the relevant authorities and maintain records and logs of the AI system's operation at the time of the incident to demonstrate compliance with the AI Act in case of ex-post audits of incidents.

While important details of the reporting framework – the time window for notification, the nature of the collected information, the accessibility of incident records, among others – are not yet fleshed out, the systematic tracking of AI incidents in the EU will become a vital source of information for improving AI safety efforts. The European Commission, for example, intends to track metrics such as the number of incidents in absolute terms, as a share of deployed applications and as a share of EU citizens affected by harm, in order to assess the effectiveness of the AI Act.

Note on Limited and Minimal Risk Systems

Finally, the limited risk category covers systems with limited potential for manipulation, which are subject to transparency obligations. This includes informing a person of their interaction with an AI system and flagging artificially generated or manipulated content. An AI system is considered to pose minimal or no risk if it does not belong in any other category.

Governing General Purpose AI

The AI Act's use-case based approach to regulation fails when confronted with the most recent innovation in AI, generative AI systems and foundation models more broadly. Because these models only recently emerged, the Commission's proposal from Spring 2021 does not contain any relevant provisions. Even the Council's approach from December 2022 relies on a fairly vague definition of 'general purpose AI' and points to future legislative adaptations (so-called Implementing Acts) for specific requirements. What is clear is that under the current proposals, open source foundation models will fall within the scope of legislation, even if their developers incur no commercial benefit from them – a move that has been criticized by the open source community <https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/> and experts <https://www.brookings.edu/articles/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/> in the media <https://techmonitor.ai/technology/ai-and-automation/ai-open-source-eu>.

According to the Council and Parliament's proposals, providers of general-purpose AI will be subject to obligations similar to those of high-risk AI systems, including model registration, risk management, data governance and documentation practices, implementing a quality management system and meeting standards pertaining to performance, safety and, possibly, resource efficiency.

In addition, the European Parliament's proposal defines specific obligations for different categories of models. First, it includes provisions about the responsibility of different actors in the AI value-chain. Providers of proprietary or 'closed' foundation models are required to share information with downstream developers to enable them to demonstrate compliance with the AI Act, or to transfer the model, data, and relevant information about the development process of the system. Secondly, providers of generative AI systems, defined as a subset of foundation models, must in addition to the requirements described above, comply with transparency obligations, demonstrate efforts to prevent the generation of illegal content and document and publish a summary of the use of copyrighted material in their training data.

Outlook

There is significant common political will around the negotiating table to move forward with regulating AI. Still, the parties will face tough debates on, among other things, the list of prohibited and high-risk AI systems and the corresponding governance requirements; how to regulate foundation models; the kind of enforcement infrastructure necessary to oversee the AI Act's implementation; and the not-so-simple question of definitions.

Importantly, the adoption of the AI Act is when the work really begins. After the AI Act is adopted, likely before June 2024, the EU and its member states will need to establish oversight structures and equip these agencies with the necessary resources to enforce the rulebook. The European Commission is further tasked with issuing a barrage of additional guidance on how to implement the Act's provisions. And the AI Act's reliance on standards awards

significant responsibility and power to European standard making bodies who determine what ‘fair enough’, ‘accurate enough’ and other facets of ‘trustworthy’ AI look like in practice.

As such, the real impact of this legislation will hinge on the EU’s dedication to implementation, the diligent execution of its provisions, the commitment to oversight, and the collaborative efforts of standard-making bodies – defining what trustworthy AI means in Europe and beyond.

Author
Mia Hoffmann <https://cset.georgetown.edu/staff/mia-hoffmann/>

Topics

Assessment Peer Watch Regions

Related Content

Understanding AI Harms: An Overview

<https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/>

August 2023

As policymakers decide how best to regulate AI, they first need to grasp the different types of harm that various AI applications might cause at the individual, national, and even societal levels. To better

understand…

Read More

Analysis

Adding Structure to AI Harm

<https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>

July 2023

Real-world harms caused by the use of AI technologies are widespread. Tracking and analyzing them improves our understanding of the variety of harms and the circumstances that lead to their occurrence once AI systems are…

Read More

Other

AI Incident Collection: An Observational Study of the Great AI Experiment

<https://cset.georgetown.edu/publication/ai-incident-collection-an-observational-study-of-the-great-ai-experiment/>

September 2023

This explainer defines criteria for effective AI Incident Collection and identifies tradeoffs between potential reporting models: mandatory, voluntary, and citizen reporting.

Read More

Data Brief

The Inigo Montoya Problem for Trustworthy AI

<https://cset.georgetown.edu/publication/the-inigo-montoya-problem-for-trustworthy-ai/>

June 2023

When the technology and policy communities use terms associated with trustworthy AI, could they be talking past one another? This paper examines the use of trustworthy AI keywords and the potential for an “Inigo Montoya…

Read More

Big Tech Is Already Lobbying to Water Down Europe’s AI Rules

<https://cset.georgetown.edu/article/big-tech-is-already-lobbying-to-water-down-europes-ai-rules/>

April 2023

CSET's Helen Toner was cited by TIME in an article about the European Union Artificial Intelligence Act.

Read More

Analysis

Agile Alliances

<https://cset.georgetown.edu/publication/agile-alliances/>

February 2020

The United States must collaborate with its allies and partners to shape the trajectory of artificial intelligence, promoting liberal democratic values and protecting against efforts to wield AI for authoritarian ends.

Read More