Examining the EU’s Artificial Intelligence Act

VB verfassungsblog.de/examining-the-eus-artificial-intelligence-act/

Paul Friedl

Gustavo Gil Gasiola

07 February 2024

Introduction

After more than two years of drafting, negotiating and renegotiating the Artificial Intelligence Act (AI Act), EU lawmakers have finally reached a consensus. With its vote on the Act on February 2nd 2024, the Committee of Permanent Representatives has at last produced a final compromise text, dispelling fears that opposition from the German, French and Italian governments could still thwart the legislative process. Whereas the Act is still awaiting approval by the European Parliament, currently envisaged for early April, barring unpredictable developments, remaining amendments will be of purely editorial nature (e.g. renumbering provisions and paragraphs in consecutive order). The academic community is thus finally in a position to provide a (slightly) more definitive evaluation of the Act’s potential to protect individuals and societies from AI systems’ harms. This blog post attempts to contribute to this discussion by illustrating and commenting on the final compromises regarding some of the most controversial and talked-about aspects of the AI Act, namely its rules on high-risk systems, its stance on General Purpose AI, and finally its system of governance and enforcement.

Researchers and civil societies initiatives have already submitted many perceptive analyses during earlier stages of the legislative process. Amongst other things, they have warned of a number of potential loopholes in the Act’s overall design and called out certain modifications as unacceptable concessions to lobbying influence. Unfortunately, as the following paragraphs will demonstrate, several consequential decisions taken on the very home stretch of the legislative process have further squandered some of the Act’s protective potential.

Duties for High-Risk Systems

Ever since its first proposal, the AI Act has adopted what is called a risk-based approach: AI systems are categorized into different classes and subject to different rules according to the level of risk they are presumed to pose for individuals and society at large. Systems deemed to pose an "unacceptable risk", such as so-called "social scoring" systems or certain types of biometric surveillance, are prohibited fully (Art. 5). Providers and deployers of "low-risk" systems, on the other hand, will only have to ensure a "sufficient level of AI literacy" on the part of those operating the system and guarantee that natural persons know when they are interacting with an AI system (Arts 4b, 52). "High-risk" systems, wedged in between these two extremes, constitute the category to which most of the AI Act's duties apply and which have therefore also garnered most of the attention over the last three years. Amongst others, providers of "high-risk" systems will need to iteratively "identify", "evaluate" and "address" their system's "reasonably foreseeable risks […] to health, safety or fundamental rights" (Art. 9), establish "appropriate data governance" practices (Art. 10) and keep logs of their system's activities (Art. 12) . The AI Act's final version now provides clarity on a number of hotly debated issues concerning the regulation of high-risk systems:

Somewhat disappointingly, the final version has adopted certain exceptions to its classification system that could allow dangerous systems to elude regulation. Prior versions of the regulation stipulated that systems will be classified as "high-risk" if they are intended to be used (for certain purposes) in one of eight high-risk areas (e.g. in the area of employment, where systems are classified as high-risk i.a. if they are used for "recruitment or selection" or "to monitor and evaluate performance"; Annex III). The Act's final version, however, now also includes a "filter provision", according to which systems will not be considered high-risk, despite falling into one of the listed high-risk areas, "if they do not pose a significant risk of harm, to the health, safety or fundamental rights of natural persons" (Art. 6(2a)). This "shall be the case" if the system is intended to (merely) "perform a narrow procedural task", "improve the result of a previously completed human activity", "detect decision-making patterns [without] replac[ing] or influen[ing] the previously completed human assessment" or "perform a preparatory task". Providers of such systems will only need to document their "reduced risk" assessment, transmit this documentation to the competent authority upon request (Art. 6(2b)) and register their system in a publicly available database (Art. 51, 60). Where the competent authority has "sufficient reasons" to believe that a provider has misclassified its system as "reduced risk", it shall investigate the system and require the provider to comply with "high-risk" duties, if it finds that the system is indeed not exempt from a high-risk classification (Art. 65a).

This final version of the "filter provision" is a considerable improvement to previous proposals, which foresaw no duty on "reduced risk" providers to register their system and would have therefore allowed them to self-classify as "reduced risk" with little to no opportunities for authorities to detect such evasion practices. Nonetheless, even the alleviated final "filter provision" provides grounds for concern: Firstly, companies will fight hard to dilate what it means for a system to e.g. "perform a preparatory task". Furthermore, in

the light of automation bias, even “preparatory” or “procedural” automation might have great influence on decision-making and material outcomes. Finally, there is reason to worry that enforcement authorities will lack resources to investigate self-proclaimed “reduced risk” systems, which might not be of high priority. Considering these aspects, the EU would have done well to adopt at least a more restrictive exemption clause, e.g. by only allowing high-risk providers to apply for a waiver without giving them the opportunity to self-exempt.

Slightly better news is provided by the institutions’ decision to include obligatory fundamental rights impact assessments (FRIAs) within the group of duties incumbent on deployers of high-risk systems, i.e. those using AI systems (Art. 29a). Early versions of the AI Act left deployers almost fully unregulated, ignoring that the way in which a system is deployed will often have a crucial impact on whether its risks materialize or not. Parliament’s version then introduced the idea of FRIAs, obliging deployers to assess and mitigate a system’s foreseeable impacts on the fundamental rights of those affected by its use – particularly marginalized and vulnerable groups. Unfortunately, however, the trilogue discussions have substantially narrowed the scope and substance of FRIA duties: Firstly, FRIAs will only need to be conducted by deployers who are either “governed by public law”, “private operators providing public services” or using an AI system for credit scoring or insurance-related risk assessment, letting the overwhelming majority of private actors off the hook. Secondly, mitigatory measures only need to be taken “in case of the materialization of […] risks”, rather than as a general prophylactic measure. Lastly, the Act’s definitive version has also undone a promising proposal by Parliament that deployers shall “involve [in their FRIA] representatives of the persons or groups of persons that are likely to be affected by the high-risk AI system, […] including but not limited to: equality bodies, consumer protection agencies, social partners and data protection”.

Next to the (albeit much too limited) inclusion of FRIAs, one may also welcome the introduction of a right for individuals to lodge complaints with competent authorities concerning possible infringements of the AI Act — even if the final version has somewhat blunted this right by deleting a corresponding duty on these authorities to respond to complaints. Ultimately, the criticism that has accompanied the legislative procedure ever since the regulation’s first draft thus by and large also applies to the AI Act’s final version: Whereas the duties on high risk system providers are numerous and extensive (if vague), considering that implementation relies centrally on providers’ self-assessment and self-enforcement (see below), it is unclear whether they will develop much practical force. Promising mechanisms that could have increased enforcement without requiring additional government resources, such as obligatory pre-market auditing by independent third parties, mandatory access for vetted researchers or more comprehensive individual remedies, unfortunately have not made the cut.

General Purpose AI Systems

The regulation of so-called General Purpose AI systems (GPAIs), often also called foundation models, presented another controversial aspect of the AI Act. Virtually unknown at the time of the Commission's original proposal, and thus also fully absent from it, GPAIs have since established themselves as a transformative new technological paradigm. Today, thousands of companies are building products on top of so-called Large Language Models, such as OpenAI's GPT, or are developing so-called generative AI, such as image or sound generation tools, which are already endangering the livelihoods of creative workers. Still other GPAIs are used to build applications that help with coding, medical imaging or different types of economic decision-making.

As is indicated by their name, GPAIs exhibit general capabilities (e.g. text analysis or image generation) that can be leveraged for many different tasks, rather than having one specific purpose or functionality. GPAIs therefore did not fit the regulation's original classificatory approach which, as mentioned above, categorizes systems as high-risk or non high-risk according to their practical purpose. Still, the need for some regulation of GPAIs, has become evident: GPAIs exhibit many of the risks that also afflict "traditional" AI systems, such as discriminatory treatment, unpredictable outcomes, subpar performance, privacy breaches, and opacity. Left unregulated, these risks could freely propagate downstream, where adopters would often lack the necessary informational, financial, and/or technological resources to mitigate them adequately.

Despite agonistic last-minute efforts by the German, Italian and French governments to prevent any regulation of foundation models, sympathizing with industry objections that warned that regulation of GPAIs would unnecessarily stifle innovation and market diversity, the AI Act's final version fortunately does extend to GPAI providers (Art. 52a-52e). Different from the Council and Parliament versions, which laid down uniform rules for all GPAIs/foundation models, however, the Act's final version now distinguishes between two different types of GPAIs systems: "conventional" GPAIs and "general purpose AI models with systemic risk". Providers of the former will only need to comply with an extremely limited set of "minimum rules" (Art. 52c): 1) Draw up technical documentation to be transmitted to the AI Office upon request (detailing i.a. the tasks the model is supposed to perform, applicable acceptable use policies, key design choices and information on the training data); 2) Draw up documentation for downstream actors intending to integrate the model (including "instructions of use" and "information on the data used for training […] where applicable"); 3) Put in place a policy to respect Union copyright law; and 4) Draw up and make publicly available a "sufficiently detailed summary about the content used for training of the general-purpose AI model", according to a template to be provided by the AI Office. Crucially, "normal" GPAI providers therefore bear no duty at all to indeed mitigate the risks created by their products, whether those concern discrimination, unreliable outputs, privacy or anything else. From both a risk-oriented and an economic perspective, a more extensive set of mandatory minimum rules for all GPAI providers would have been much preferable.

This reluctance to substantially regulate “normal” GPAIs makes the obligations reserved for “systemic risk” GPAIs all the more important. “Systemic risk” GPAIs, in addition to these transparency obligations, need to “perform model evaluation” (although it remains somewhat unclear in what regard), report to the AI Office “serious incidents and possible corrective measures to address them”, “ensure an adequate level of cybersecurity protection” and “assess and mitigate […] systemic risk at Union level”, i.e. risks that can have “significant impact on the internal market”, because they have at least “reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole” (Art. 52d). Regarding classification, Article 52a establishes that a model should only be considered “systemic risk” where it either a) has “high impact capabilities”, i.e. “capabilities that match or exceed the capabilities recorded in the most advanced general purpose AI models”, or where b) the Commission has taken a decision to classify a specific system as “systemic risk”. Giving the distinction an inadequate techno-centric twist, paragraph 2 then establishes that a system shall be presumed to have “high impact capabilities”, if the amount of compute used for its training was greater than 10^25 FLOPs. For comparison, Open AI’s GPT-4, considered by some the most powerful foundation model currently available, is presumed to have required around 10^25 FLOPs. By today’s (!) numbers, presumably less than a handful of models would thus be classified as “high impact capabilities”, while there are of course many very capable models available on the market. These criteria thus, not only seem to focus on an ultimately secondary dimension, i.e. technical specifications (rather than social impact), but furthermore seem extremely laissez-faire. Although the Commission “shall amend” these classificatory thresholds “in light of evolving technological developments, such as algorithmic improvements or increased hardware efficiency”, the provision thus leaves the impression that substantial rules will be limited to an extremely limited number of mega-actors. FLOPs, one might say, equal the VLOPs.

Finally, it is debatable whether the AI Act has found the best compromise regarding the regulation of Open Source GPAIs. Open Source GPAIs exhibit a number of significant benefits: they catalyze the distribution of technology and know-how necessary for smaller players to be able to challenge reigning incumbents and thus stimulate innovation. They also provide a level of transparency largely unknown from commercial models (e.g. regarding training data, model weights or compilation code), which is crucial for researchers to explore the risks of foundation models and also allows for greater levels of safety, as users can inspect a system’s behavior much more thoroughly. The AI Act now stipulates that, except for the duties to respect copyright law and to provide a summary of the utilized training data, non-”systemic risk” models “that are made accessible to the public under a free and open licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available,” are exempt from compliance with the Act’s GPAI regime (Art. 52c(-2)). The EU thus takes the position that the risks of GPAIs are sufficiently offset by the higher transparency provided by Open Source models. In light of the current trend of “open washing”, where big companies label their models “Open Source”,

mostly to exploit and ultimately re-privatize the labor of unpaid developer communities, knowing that due to their immense resource-intensity alone the released models can hardly be used or reproduced by anyone else, this decision seems problematic. It also does not sufficiently account for legitimate concerns that the uncontrolled release and distribution of GPAIs, distinctive of Open Source projects, could allow powerful models to get into the hands of “bad actors” (e.g. disinformation actors) and that this should be prevented by retaining avenues for prohibiting the open-sourcing of certain models. Again, the compromise seems to tilt quite heavily in favor of industry interests and ultimately lacks an adequate balancing of competing concerns.

System of Governance

Whether the AI Act will prove effective is a question that extends beyond its substantive duties; the governance system established by the regulation will ultimately be at least equally pivotal. It is therefore crucial to have a basic understanding of the different institutions cooperating for the Act’s application and implementation. Next to the Commission, which is entrusted with a great many of tasks, mostly to be executed through delegated acts (see below), there are essentially five other important institutions mandated with implementing and enforcing the Act: 1) The AI Office: Envisaged by the Parliament as a strong, independent body, the final version reduces the AI Office to an agency attached to the Commission, raising concerns whether the it will dispose of sufficient financial resources (Art. 55b). Next to the important task of overseeing compliance by GPAIs, the AI Office will i.a. also facilitate the drawing up of Codes of Practice (Art. 52e; see below); 2) The European Artificial Intelligence Board (“Board”): Composed of one representative designated by each Member State (Art. 56), the Board shall advise and assist the Commission and the Member States in order to facilitate the consistent and effective application of the Act. This includes duties such as the issuing of “recommendations […] on any relevant matters related to the implementation of this Regulation” or contributing to the “harmonisation of administrative practices in the Member States” (Art. 58); 3) The Advisory Forum: Intended as another advisory body for both the Commission and the Board, the Advisory Forum ”shall represent a balanced selection of stakeholders, including industry, start-ups, SMEs, civil society and academia”, to be appointed by the Commission (Art. 58a). Consultation of the Forum is obligatory only for the drafting of standardization requests and common specifications (see below); 4) The Scientific Panel of Independent Experts: As another advisory body, the AI Act’s final version furthermore establishes the so-called Scientific Panel, to be composed of independent, Commission-appointed technical experts (Art. 58b). Next to assisting the enforcement activities of other bodies, the Scientific Panel also plays a key role in the enforcement of rules for GPAIs, as it may launch qualified alerts to the AI Office where it suspects that a GPAI poses “systemic risk”. 5) National supervisory authorities: Finally, at the Member State level, the AI Act above all requires the designation of at least one so-called market surveillance authority (Art. 59). The AI Act generally does not oblige high-risk system providers to undergo authorized third-party assessment or obtain some form of official

certification (the only exception from this being systems using biometric technology, Art. 43(1)). Rather, conformity with the Act's duties will generally only be assessed by providers themselves (so-called "conformity assessment procedure based on internal control", Art. 43(2)). The main responsibility to supervise the Act's implementation therefore lies with national market surveillance authorities, whom Member States need to endow with all authorities necessary for "market surveillance, investigation and enforcement" of the AI Act (Art. 3(26)). This will include amongst other things the power to start ex officio investigations, the power to demand operators to provide all necessary information for evaluation a system's conformity and the power to demand operators to take all necessary corrective actions to end non-compliance (Art. 63, Regulation 2019/1020).

These governance bodies of the AI Act are not only tasked with overseeing and enforcing compliance, but will also play a crucial role in specifying, amending, and/or derogating certain provisions. Indeed, almost no provision of the Act does without any such authorization allowing for the law's modification, whether those be mandatory or optional. Many of these supplementary decisions will need to be taken by the Commission. For instance, within 18 months the Commission will need to "provide guidelines specifying the practical implementation" of the above-mentioned provision allowing for an exceptional classification as "reduced risk" (Art. 6(2c)). "When necessary", it shall adopt delegated acts amending the above-mentioned thresholds determining if a GPAI presents "systemic risk" (Art. 52a(3)). Some delegations, however, fall upon other bodies, such as the AI Office or the Scientific Panel. So-called harmonised standards, codes of practice and common specifications, however, will arguably play the most crucial role in concretizing the Act, as providers which are in conformity with any of these types of delegated "legislation" will benefit from a presumption of conformity regarding the Act's respective requirements themselves (Arts 40-41, 52e). Standards will be developed by European standardisation organisations (ESOs), which constitute a rather troubling choice: ESOs are technical institutions and thus rather unqualified to decide upon the many value-laden questions that arise under the AI Act's obligations. What is more, existing standardization processes are highly opaque and by and large controlled by industry representatives. Codes of practice, which are to concretize the duties of GPAI providers, will be drawn up by the AI Office. The provision governing the development of codes of practice, however, stipulates that the AI Office shall invite GPAI providers to participate in the development process, while limiting other "relevant stakeholders" to "supporting the process", thus again indicating a prioritization of industry interests. Finally, where one of the Commission's standardization requests has not been fulfilled or where a Code of Practice has not been finalized by the time the Regulation becomes applicable, the Commission can also opt to specify the Act's duties itself by adopting common specifications in the form of implementing acts. The decision to rely on delegated rule-making to specify the Act's many underdetermined obligations is not blameworthy per se. It introduces much-needed normative flexibility to a regulatory field that

is still quickly developing. At the same time, however, the AI Act does too little to ensure that these supplementary legislative processes include broad, representative groups of diverse stakeholders and risks privileging regulated parties.

Conclusion

Whether the AI Act will deliver on its promise of protecting individuals and societies from harm while simultaneously boosting European digital economies – the by now well-known, slightly schizophrenic dyad of EU regulatory objectives – is of course still unclear. Much will depend on the concrete shape of the many implementing acts and the secondary legislation yet to be passed. However, as the preceding paragraphs were hopefully able to show, many of the decisions taken in the very final phase of the legislative process have rather weakened its prospects. While it still does not seem impossible to render the AI Act into an effective tool, safeguarding citizens vis-a-vis risky AI applications, this will require persistent efforts by well-resourced administrative bodies making many right decisions. In light of these considerations, one may hope that the AI Act will function as a springboard, rather than a ceiling for the global regulatory efforts currently still ramping up.

SUGGESTED CITATION Friedl, Paul; Gasiola, Gustavo Gil: *Examining the EU's Artificial Intelligence Act, VerfBlog,* 2024/2/07, https://verfassungsblog.de/examining-the-eus-artificial-intelligence-act/, DOI: 10.59704/789d6ad759d0a40b.