Statement on the AI Act Trilogue Results

Philipp Hacker[*]

This version: December 10, 2023

The conclusion of the AI Act Trilogue has been a significant moment for the AI community and the broader AI environment. Congrats to and deep respect for the lawmakers who stayed strong during this this marathon session!

In this initial reaction, I am providing an overall assessment, as well as more specific commentary on the AI Act deal of December 8, including a critique of several aspects (e.g., AI definition; foundation models/GPAI; open source treatment).

I. Overall assessment

The political agreement on the AI Act is extremely important in a dual sense. First, it sends a strong signal that the EU is still functional as a major force in international technology regulation.

Second, in many respects, the contents strike a sensible balance between allowing innovation and protecting fundamental rights and public safety. But some gaps remain. The rules for foundation models (FMs) are a step in the right direction, but do not go far enough. The minimum standards are actually extremely toothless (mere transparency, copyright) – a tiger too toothless, in my view. Even 10^24 FLOP models exhibit AI safety and cybersecurity risks that cannot be left to self-regulation. **If you want to play Champions League, you have to stick to the Champions League rules.**

This is why FM regulation is necessary: If you exclude the FMs, the regulatory burden is shifted to the downstream providers. Fixing the error in the deployment a thousand times is worse than tackling the problem at the source (= FM), a clear least-cost avoider argument from standard (and very economically liberal) law and economics. FM regulation is efficient, self-regulation is inefficient and dangerous in this domain.

Does sensible FM regulation deter innovation? No. A new study finds that even for quite advanced but not even top-notch 10^24 FLOPs models, such as Bard, ChatGPT etc. (i.e., lower than GPT-4 and Gemini), expected compliance costs only add up to roughly 1% of total development costs (https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf). This is a sum that everyone, including Mistral, Aleph Alpha etc., can and should invest in basic industry best practices for AI safety.

Third, however, the attractiveness of the EU as a future hub for AI innovation and deployment should have been strengthened: the **AI Act deal should have been paired with an announcement of massive amounts – in the dimension of billions of euros** – in EU and collective Member State funding for AI research and deployment: in compute, chips infrastructure, and talent retention. Only in this way, we can secure strategic independence in a key technology of the 21st century, and prevent the same geostrategic dependencies that

[*] Professor Dr. Philipp Hacker, LL.M. (Yale), Research Chair for Law and Ethics of the Digital Society, European New School of Digital Studies, European University Viadrina Frankfurt (Oder); co-lead, RECSAI.

brought Europe to the border of chaos in the field of oil and gas supply. Europe is lagging far behind when it comes to cutting-edge AI model production – with only very few exceptions –, and this is clearly becoming a geostrategic problem in the current international environment. Inter alia, we need a well-funded European DARPA.

II. Commendable parts of the AI Act deal

The recent advancements in the AI Act Trilogue have resulted in several commendable elements.

1. Alignment with existing sectoral regulation

Among these, the alignment with existing sectoral regulation stands out as particularly crucial. This alignment ensures that the AI Act does not operate in isolation but rather complements and integrates with the broader regulatory framework. It is essential, however, to remain vigilant, and wait for the final text, to ensure that this alignment is comprehensive and does not leave critical areas doubly, inefficiently, or even conflicting really regulated (such as medical AI, credit scoring, and insurance).

2. Research exemption

Second, the exemption for research within the AI Act is significant and laudable. Nonetheless, the boundaries of this exemption appear somewhat ambiguous, especially when considering the common academic practice of publishing research outputs (including models) in publicly accessible repositories. The spirit of academic research is grounded in the principles of openness and verifiability. When research findings, particularly those related to AI, are published in open repositories, they enable external verification. The AI Act must clearly delineate how it accommodates such practices without imposing undue restrictions on the dissemination of knowledge while also ensuring that the publication of research does not inadvertently circumvent the Act's safeguards designed to protect the public and uphold AI safety standards.

3. Foundation Models (GPAI) Regulation

a) Minimum standards for all FMs

The Trilogue has led to the establishment of minimum standards for all foundation models, now referred to as general-purpose AI (GPAI) models. These standards revolve around transparency, including watermarking, and adherence to copyright provisions. This is crucial to provide for accountability, but falls short of more meaningful rules that the EP had foreseen in its position: **rules on cybersecurity, content moderation, and AI safety would be crucial for all FMs** (see below, III.).

b) Provisions for Systemic Risk FMs:

Additional obligations for systemic-risk FMs have been rightfully introduced:

- **Risk Management**: Organizations must perform model evaluations using state-of-the-art protocols and tools.

- **Red Teaming**: There is a necessity to conduct and document adversarial testing to identify and mitigate systemic risks.
- **Cybersecurity**: Maintaining an adequate level of cybersecurity for both the AI model and its physical infrastructure is non-negotiable.
- **Energy Consumption**: Entities must track, document, and report on the known or estimated energy consumption of the model.

This is fairly comprehensive and good. But the threshold of 10^25 FLOPs for a default categorization of systemic risk models is too high. Currently, to my knowledge, only GPT-4, potentially Gemini, and perhaps one or two other models, surpass this threshold (see the very useful study by The Future Society: https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf).

Lowering this to 10^24 FLOPs would be more inclusive of large models that already demonstrate systemic risks and are currently on the market (GPT-3.5; Claude; Bard).

c) Copyright Provisions

The copyright provisions within the AI Act are reasonably constructed.

- **Copyright Compliance Regime**: Providers must implement a policy that respects Union copyright law, utilizing state-of-the-art technologies where appropriate. This is tantamount to a compliance regime, putting in place organizational and technical measures to ensure heating the opt-out rights of rightholders. This makes sense as only companies systematically violating copyright provisions would not install such a system.
- **Training Content Summary**: Providers are also required to draw up a sufficiently detailed summary of the content used for training the AI model.

Critique of Copyright Provisions:

- **Detail of Summary**: It must be made explicit that the summary does not need to delve into individual training data points, which would be prohibitively expensive. This may be clarified in the Recitals (this may actually have been included in the final text of the Dec 8 deal).

4. Regulated Self-Regulation and Safe Harbors

Code of Practice: The Commission's power to approve codes of practice for FMs will endow the Code with general validity within the Union. This is an excellent development. It presents an opportunity to leverage decentralized industry and expert knowledge and operationalize vague concepts for specific sectors. This helps to establish safe harbors for companies. These will be crucial to attract and retain companies, and talent, in the EU.

5. AI Value Chain

The information sharing along the AI value chain and the status of deployers only fine-tuning FMs are noteworthy aspects of the regulation.

Importance of the AI Value Chain:

- **Information Sharing**: The sharing of information is crucial for maintaining transparency and accountability, and for making sure that all actors have the required information for AI Act compliance.
- **Fine-Tuning Exemption**: It is crucial to ensure, via a presumption, that deployers who are only fine-tuning FMs should not be subject to GPAI rules. We will have to wait for the final deal text to see if this is included.

6. Environmental impact and sustainability

The AI Act's inclusion of provisions concerning the environmental impact of AI systems is a commendable step toward sustainable AI regulation. Although these provisions represent a foundational recognition of the importance of environmental considerations in AI, they fall short of a more comprehensive framework (see my paper "Sustainable AI Regulation"). It is vital that future iterations of the AI Act expand upon these rules to ensure that the AI industry progresses in an environmentally sustainable manner, with a more rigorous approach to the assessment, mitigation, and ongoing management of the environmental footprint of AI systems (sustainability impact assessments; consideration of an extension of the Emissions Trading System to data centers and other high-consuming IT processes).

7. Subjective rights for citizens

Another good aspect of the AI Act is its facilitation of citizen complaints and the incorporation of explanation rights. This empowers individuals, fostering a transparent AI ecosystem where citizens can seek redress and understand how AI decisions are made that affect their lives, beyond the arcane provisions of Art. 22 GDPR (see the recent Schufa judgment of the CJEU and the Amsterdam cases in the Uber/Ola proceedings). These features are critical for building trust in AI technologies and ensuring that AI operators remain accountable.

III. Critique

1. OECD Definition of AI

Revised OECD Definition: “An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.”

In this unqualified form, this is a definition of software, not of AI. Take an auto-sum function in an excel sheet. It has an objective (building a sum), input (entries), and an output that may influence environments (as per the relevance of the sum for any decisions). So, the only distinguishing criterion for AI on this definition is “infers”. I sincerely hope that the Recitals will make clear that it takes 1) independence from human intervention (=automation) AND 2) significant adaptability/learning capacity to qualify for “AI-relevant inference”. Otherwise, the AI Act will be a Software Act – and that would have to be construed quite differently.

2. Critique of GPAI Regulation for all Foundation Models

The following provisions are missing from the minimum requirements for all FM providers:

- **Cybersecurity**: It is imperative to ensure an adequate level of cybersecurity protection for all foundation models, particularly given the current geopolitical situation with increasing threats and wars. Insufficient cybersecurity measures propagate down the AI value chain and may open backdoors to a wide variety of applications for malicious actors, with a state or non-state background.

- **Content Moderation**: Expanded content moderation is necessary to mitigate the proliferation of hate speech and fake news. The lack of rules concerning robust content moderation measures is concerning. Experiments have shown that foundation models are prone to provide illegal outputs, including hate speech. Making sure that this potential cannot be abused by malicious actors seems paramount. Furthermore, content moderation must ensure that foundation models behave appropriately particularly also in dealing with "advice" provided by foundation models concerning physical or mental health problems. Overall, if the FLOPs threshold defining systemic risk models remains at 10^25, most powerful GPAI models will only have to meet the minimum standards. This oversight can be rectified by mandating, for all FMs, a compliance system to prevent illegal outputs, feasible for companies of varying scales, including Aleph Alpha and Mistral. The compliance system should ensure, via state-of-the-art technical and organizational measures, that content generated by AI, whether audio, image, video, or text, abides by the laws of Member States from which the model is accessible.
    - **Extension of DSA Provisions**: The provisions of Articles 16 and following of the Digital Services Act, including trusted flaggers and a notice-and-action mechanism, should urgently be extended to the domain of Generative AI (see our paper https://dl.acm.org/doi/abs/10.1145/3593013.3594067). The reason for this is to establish a more effective and decentralized system for flagging and removing illegal content generated by AI systems to stem the tide of hate speech and hallucinations still plaguing GenAI – crucial ahead of the next global election cycles (US, EU, and beyond). This mechanism would bolster the existing content moderation framework by incorporating community-driven oversight. It would ensure a broader base for monitoring and mandate a quick response to violations highlighted by trusted flaggers (e.g., registered NGOs).

- **AI Safety**: There is an urgent need for comprehensive strategies to mitigate risks associated with cyber malware and biochemical terrorism. Again, even smaller foundation models may pose significant risks here. Mandatory provisions for all FM providers are essential. This includes
    - Mandatory red teaming for all FMs.

3. Access for Vetted Researchers:

- Vetted researchers should have the right to access foundation models, akin to Article 40 of the Digital Services Act (DSA). The rationale for this is to allow for independent verification of stress tests and benchmarks, as well as decentralized monitoring. It's great that companies are doing research on this, but all results must be verified externally – that's just standard academic practice. Such access ensures that oversight does not solely rest with the providers of the models alone (and notoriously resource-constrained regulatory bodies) but involves the academic community at large.

4. Systemic Risk Foundation Models

Critique of Systemic Risk FMs/GPAI:

- **FLOPs Threshold**: As said above: The threshold of 10^25 FLOPs for a default categorization of systemic risk models is too high. Currently, to my knowledge, only GPT-4, potentially Gemini, and perhaps one or two other models, surpass this threshold (see the very useful study by The Future Society: https://thefuturesociety.org/wp-content/uploads/2023/12/EU-AI-Act-Compliance-Analysis.pdf).
  - **Important:** Lowering this to 10^24 FLOPs would be more inclusive of large models that already demonstrate systemic risks and are currently on the market (GPT-3.5; Claude; Bard). This is what the Commission should do via a Delegated Act.
- **External Red Teaming**: While the rules are sound, red teaming would benefit from the involvement of external entities to ensure an unbiased and comprehensive assessment.

5. Open Source (OS) Exemptions

The Trilogue exempts pre-trained AI models made accessible under an open-source license from the minimum standards (if I interpret Art. C(4) vs. Art. 2 correctly). Open-source models are not excluded from the provisions on systemic risk GPAI.

Critique of OS Exemptions:

- **OS Models Threshold**: The current exemption for powerful OS models up until a 10^25 FLOPs threshold is questionable. A lower threshold, including minimum standards for 10^24 FLOPs OS models, would ensure that such models are regulated appropriately.

- **OS Prohibition:** For quite highly-performing OS models, such as 10^23 FLOPs model, I would even recommend a prohibition of open sourcing. These models should only be made available in a hosted access model. Studies show that safety layers can be easily removed once a model can be fully downloaded. These models are essentially dual-use goods, and cannot be made freely available for any use and modification for the general public. Otherwise, we have significant public safety threats, including cyber malware, and bio- and chemical terrorism.

6. Investment in AI and AI Safety

Investment in AI infrastructure and safety mechanisms is a pivotal aspect of the EU's strategic direction.

Necessity for Investment:

- **EU DARPA**: There is a need for an EU equivalent to DARPA to ensure digital sovereignty and substantial investment in AI infrastructure.

- **AI CERT**: The establishment of an EU-level AI CERT is critical for addressing public safety threats efficiently; see, in this direction, the proposals made by Ramayya Krishnan and Martial Hebert (https://thehill.com/opinion/technology/4079196-back-to-the-future-look-to-the-1980s-for-guidance-on-ai-management/).

In light of these observations, the Trilogue's results, while a leap in the right direction, present clear opportunities for enhancement. Ensuring a robust and comprehensive regulatory framework for AI is essential for a future-proof digital Europe.