What’s Missing from the EU AI Act

Philipp Hacker 2023-12-13T20:34:47

The AI Act negotiators may still have been recovering from the political deal that was struck during the night of December 8 to 9 when two days later Mistral AI, the French startup, open sourced its potent new large language model, Mixtral 8x7B. Though much smaller in size, it rivals and even surpasses GPT 3.5 on many benchmarks thanks to a cunning architecture combining eight different expert models. While a notable technical feat, this new release epitomizes the most pressing challenges in AI policy today, and starkly highlights the gaps left unaddressed by the AI Act: mandatory basic AI safety standards; the conundrum of open-source models; the environmental impact of AI; and the need to accompany the AI Act with far more substantial public investment in AI.

Why We Must Address These Issues Now

These challenges are not just theoretical concerns but real and immediate. The rapid advancement in AI technologies, exemplified by the recent releases of Google’s Gemini, Mixtral 8x7B, Claude 2.1 et al., requires an equally swift and thoughtful policy response. The current AI Act, while a step in the right direction, does not sufficiently tackle core issues, leaving the EU vulnerable in key areas of AI development and regulation.

After three days of intense negotiations, the EU did converge on minimum standards for all foundation models (called general-purpose AI models in the AI Act) and more stringent rules for so-called high-impact foundation models with systemic risk. However, the minimum standards are actually extremely weak – a tiger too toothless, in my view. They include mere transparency and limited copyright provisions. By default, stricter rules for high-impact models kick in if the model was trained with more than 10^25 FLOPs (floating-point operations, roughly equivalent to calculation steps). However, for as much as we know, only GPT-4, and perhaps Gemini as well as one or two other models, cross that threshold. As the recent Mistral model shows, the tendency is to develop more potent smaller models. Yet, even “smaller” models, for example in the range of 10^24 FLOPs (e.g., Bard, ChatGPT), exhibit significant 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. And those rules are binding, not voluntary like the Disinformation Code of Practice that Twitter famously withdrew from when its new owner did not like it anymore**.**

Such a framework is clearly insufficient when it comes to the currently most powerful technology. If regulation excludes foundation models (FMs), the regulatory burden is shifted to the downstream deployers. Fixing the error in the deployment a thousand times is worse than tackling the problem once at the source (= FM) – a clear least-cost avoider argument from standard (and very economically liberal) law and

economics. Rightly tailored foundation model regulation is economically efficient, and self-regulation is both inefficient and dangerous in this domain. Against this background, here is what needs to be done to plug these gaps in the AI Act going forward:

Mandatory Basic AI Safety Standards

The first glaring omission in the AI Act is a comprehensive framework for AI safety for all foundation models, including cybersecurity, mandatory red teaming against illegal content, and content moderation. Large language models are capable of generating content at an unprecedented scale. This makes them highly valuable assets in many high-impact domains, from medicine to education. However, without stringent guardrails, this opens the floodgates not only to a deluge of potential misinformation and hate speech, but also cyber malware, and help for biological and chemical terrorism, as the Dutch Cyber Security Center notes. 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. To counteract this, best industry practices already include red teaming and the introduction of safety layers to guard against such abuse by malicious actors.

The AI Act should have mandated this for *all* foundation models, but it did not. A way forward is to require a robust, decentralized content moderation system, much like the Digital Services Act (DSA). If the AI Act does not take this on board in the next few weeks, it could be added to a revised DSA. 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. The reason for this is to establish a more effective and decentralized system for flagging and removing toxic, harmful, or outright dangerous content generated by AI systems still plaguing GenAI – crucial ahead of the next global election cycles (US, EU, and beyond). This mechanism would bolster the existing, but voluntary industry practices by incorporating community-driven oversight (e.g., via registered NGOs).

Some might argue that these measures could stifle innovation or are too ambitious. However, the rapid development and potential risks of AI technologies necessitate bold steps. Does sensible FM regulation deter innovation? The plain answer is: 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. This is a sum that everyone, including smaller European providers such as Mistral, and Aleph Alpha, can and should invest in basic industry best practices for AI safety.

Balancing Open-Source Innovation and Public Safety

The decision to release Mixtral 8x7B as open-source, just like Meta's Llama 2 or the Falcon family, while championing transparency and accessibility, highlights

significant public safety concerns. Generally, open-source models present undeniable advantages that are essential in the broader AI landscape. They act as a counterbalance to monopolizing tendencies in the foundation model market, fostering a more diverse, competitive, and accessible AI ecosystem. However, once powerful enough, the risks of open sourcing arguably outweigh the benefits. Unregulated access to such powerful models can lead to malicious abuse, including malware generation and terrorist uses. Importantly, if the model can be downloaded, safety layers can be quite easily – and even inadvertently – removed. Hence, the EU seriously needs to rethink its stance on open-source AI models, which are currently freed from regulation unless they constitute systemic-risk models. From a certain performance threshold on (e.g., well below the current one: e.g., 10^23 FLOPs or GPT 3.5 equivalent benchmark performance), a prohibition on full open sourcing should kick in, at least until we better understand how to engineer safety features into the models. Rather, a more controlled access system, where usage can be monitored and regulated via hosted access, is necessary. Importantly, this doesn't mean stifling innovation but rather channeling it responsibly.

Such rules should be coupled with a framework granting access to vetted researchers – Article 40 DSA provides a template. The rationale again is to allow for independent verification of stress tests and benchmarks. Thus, open models would be closed, in a responsible way (via hosted access), but closed models would simultaneously opened up (via vetted researchers). Even OpenAI would then have to live up to its name again. Currently, it is great that many companies are doing voluntary safety testing and research, but these results must be verified externally – that's just standard academic and safety 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. Trust, but verify.

Addressing AI's Environmental Footprint

Large AI models relate to another, truly existential "safety concern": climate change. The AI Act deal includes the first provisions concerning the environmental impact of AI systems, a commendable step toward sustainable AI regulation. However, they fall short of a more comprehensive framework. While AI applications can be beneficial for the environment, the astronomical computational power and water resources needed for training and deploying large-scale AI models also significantly contribute to climate change, complicating global sustainability efforts at a time when immediate action is vital. By 2027, AI models are projected to consume the energy equivalent of a country like Argentina or the Netherlands. 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 – including mandatory sustainability impact assessments and a possible extension of the Emissions Trading System to data centers and other high-consuming IT processes.

Boosting Investment in AI

Finally, and most importantly from an economic perspective, the European AI framework overlooks a critical component for advancing AI innovation in Europe, and beyond – public investment. Norway alone has just dedicated 1 bio. NOK to AI advancement; the UK invests 300 Mio. pounds in AI supercomputing. To compete on a global stage with AI powerhouses like the US and China, the EU must significantly increase, pool, and make visible its public investment in AI. The aim should be to not only match but surpass current AI capabilities, as evident in models like GPT-4, and to pave the way for sustainable AI technologies. This is not an issue for single Member States.

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 infrastructure, chips production, and talent retention. Only in this way, the EU can secure strategic independence in a key technology of the 21st century, and prevent the same geostrategic dependencies that 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.

Overall, the December deal on the EU AI Act is a commendable starting point, but it's time to build upon it. Now, the hard work starts. Not only in ironing out the technical details, but also in setting the right priorities, between ensuring minimum standards in critical safety areas and fostering a climate for innovation and development. This necessitates a collaborative effort from policymakers, industry leaders, and the academic community to revisit and refine the Act. We must not only meet the challenges posed by AI but harness its full potential for a progressive, competitive, and responsible digital Europe.
