Microsoft

13 October 2023

EU Artificial Intelligence Act:
Microsoft trilogue recommendations

In view of the ongoing EU AI Act negotiations, in particular with regard to regulation of foundation models and general purpose AI systems, we highlight below several key recommendations for consideration by the co-legislators, building upon our earlier recommendations.

Summary of our recommendations

Apply calibrated and risk-based requirements to both **advanced foundation models** and **advanced general purpose AI applications** (in addition to requirements for high-risk systems)

- ***Model-level requirements are appropriate*** as part of a durable and comprehensive regulatory framework, though they should be placed only on *advanced* foundation models and calibrated to the types of risks providers of such models are able to address. A threshold based on amount of compute used in training should be used to determine when a model is considered advanced, and should be set and adjusted over time with close involvement by the AI Office.
- ***Application-level requirements are appropriate*** for *advanced* general purpose AI apps that are highly capable across a range of functions and intended to be deployed in a wide range of use cases given the (moderate) risks these systems can pose if not developed, deployed, and used responsibly.
- ***A focus on advanced technologies*** preserves the Act's risk-based approach as it extends from high-risk systems (use cases) to models and general purpose applications (technologies).

ADVANCED FOUNDATION MODELS

1. Clearly define 'advanced foundation models' and apply calibrated, risk-based requirements

Focus on advanced foundation models: The AI Act should define *advanced foundation models,* focusing only on the most capable category of models on the market and anticipated and applying to them requirements for risk mitigation and technical documentation.

- A "foundation model" should be defined as an AI model that is trained on broad data at scale, is designed for generality of output, and is *intended to be adapted and integrated into a variety of downstream applications* to *complete a wide range of distinctive tasks*.
- An "advanced foundation model" should be defined as a foundation model *trained on a very large amount of compute above a high threshold*. This threshold should be set and adjusted over time by the Commission and AI Office in consultation with industry and civil society experts. A compute threshold based on Floating Point Operations Per Second (FLOPs) on which a model was trained should be utilized given the consistent connection between amount of compute used to train a model and the resulting capabilities of the model (versus other measurements, such as model size as determined by number of parameters). A compute threshold is also available to reference *ex ante*, as an indicator of potential risk.

Calibrate requirements to model provider responsibilities: Model providers, including developers and those that deploy or provide access to models through platform or API services, should be required to focus on risks within their control – as distinct from risks more effectively addressed by downstream application developers. Requirements should thus be limited to:

- Risk assessment and mitigation of identified risks
- Red teaming by a team independent to the product team
- Development of the model to perform robustly when relevant requirements applicable at the platform and/or application level are also applied
- Provision of technical documentation and intelligible instructions for use to help enable downstream providers to comply with their obligations:
  - Name and address of provider
  - Description of model capabilities and limitations and identified risks
  - Description of the model's performance
  - Member States in which the model is placed on market
  - Further technical documentation as agreed by contract to address the varied nature of the AI value chain
- Development of a management system to oversee compliance with the Act

Foundation model providers should not be required to: 1) publish a list of copyrighted training data, which could undermine the EU's text and data mining exception, impact IP protections, and provide insights for malicious actors; 2) meet Art. 52 transparency requirements that better apply to applications; or 3) subject models to misleading, duplicative, or inconsistent third party audits, instead fostering ecosystem readiness (see section 5) and allowing use of suitable alternatives in the interim, such as self-attestation and documented evidence.

ADVANCED GENERAL PURPOSE AI APPLICATIONS

2. Place proportionate requirements on advanced general purpose AI applications

Focus on advanced general purpose AI (GPAI) applications: Where it concerns general purpose AI systems, the Act should be clearly scoped around *applications*, complementing separate requirements for advanced foundation *models*, and avoid sweeping in lower risk application-level GPAI systems that perform general functions (e.g. a conversational interface that facilitates more effective use of a software product).

- "Advanced general-purpose AI system" should be defined as an application-level system that 1) performs a *wide range of functions*, such as generating code, interpreting images, answering questions, providing advice, and creating complex content; 2) is intended to be used across a *wide range of different use cases and domains*; and 3) may pose a *meaningful risk of harm if not used in a manner consistent with a provider's guidance*.

Calibrate requirements to responsibilities of advanced GPAI application providers: Just as model providers are well placed to address certain risks, so are application providers well placed to address others. Advanced GPAI application providers should:

- Perform a risk assessment and mitigate identified risks
- Provide a plain language description of the capabilities and limitations of the application, the factors that will affect its use, and use cases for which it is not suited
- Notify a user that it is interacting with AI
- Notify a user that audio and visual content produced by the application is AI generated.

Maintain Annex III approach to high-risk systems: If an entity deploys a GPAI application (advanced or not) in an Annex III use case, then it should be designated a high-risk provider for that system. Prior to such a deployment, the high-risk provider should request technical information to meet high-risk system requirements as appropriate, with the GPAI application provider free to decline provision of information if such a use is prohibited in its terms of service.

OTHER MODELS AND APPLICATIONS

3. Exclude less advanced models and GPAI applications from the Act's scope or limit requirements to only basic and applicable transparency steps

Less advanced foundation models, and the broad ecosystem of their developers, should not attract requirements as such models pose more limited risks that can be adequately addressed at the application layer. Less advanced models may be developed to be simpler alternatives to more powerful models as a cost-effective option for more basic tasks. There is a large and growing ecosystem of AI models (over 343,000 models on Hugging Face alone). Many of these are and will be open source models that help spread the benefits of AI and advance innovation and an understanding of AI risk. To the extent that less advanced models attract any requirements, they should be limited to basic transparency regarding model capabilities and limitations.

Less advanced or lower-risk GPAI applications should likewise be out of scope or only subject to basic transparency requirements, such as the Act's requirements for transparency around when audio and visual content is AI generated and to notify users that they are interacting with an AI system when that is not clear from the context of use.

AI GENERATED CONTENT

4. Ensure content transparency requirements apply only to generative audio and visual content

Utilize state of the art tooling: Providers of applications that create audio or visual content should utilize "state of the art" provenance tooling so users know when content has been AI generated and if and how it has been changed since creation. While the "state of the art" will develop over time, current tooling built on standards such as C2PA offers significant promise.

Focus on audio and visual content: Requirements should apply to audio or visual content but not text. Given the visceral power of audio and visual content, the risk that it's used deceptively without being distinguished as AI generated, as in the context of "deep fakes," transcends that associated with text. Moreover, the way AI tools power text generation, and the limitations to provenance tooling currently available mean that requiring content transparency for text could ultimately mislead users. While leading provenance tools are applied as the final signing step for a complete file before content is shared or published, current text-based generative AI products and services are rarely used to generate complete documents. Instead, they are used in conversational turns to generate sentences or sentence fragments, and the resulting text is usually modified and validated by a human to various degrees and moved from one file format to another. Efforts to embed watermarks in text, inserting data with hard-to-detect modifications, often impact the quality of generated text, and if the text is edited, then the mark degrades.

AI OFFICE

5. Establish an AI Office that can collaborate with national authorities and international partners to future proof the Act and advance work on evaluations

Ensure the Act remains future proofed: As technology and our understanding of risks and risk mitigation techniques evolve, elements of the Act, such as the compute thresholds used to identify advanced foundation models, may also need to be adjusted.

Collaborate with external experts and international partners: The proposed Advisory Forum could help facilitate shared learning and coordination with diverse multi-stakeholder experts.

Advance work on evaluations, coordinated enforcement, and mutual recognition: Effective evaluation techniques must be developed to assess risk mitigations for advanced foundation models. As evaluation techniques are defined, an ecosystem of assessors will continue to emerge and increase in maturity and capacity. To ensure consistency in assessments that can be re-used across the EU's Digital Single Market, expertise and processes must be established to certify assessors, including independent third-party assessment organizations, and manage oversight, including of assessors' disputed interpretation of requirements or evaluation findings. The AI Office should manage or support these functions. Article 58a on benchmarking in the Parliament text should be adopted, with the AI Office tasked to collaborate with national authorities and international partners to develop 1) guidance for measuring and benchmarking AI systems, particularly advanced foundation models, and 2) criteria for certifying national or third-party assessor readiness to ensure consistency and enable re-use of assessments across jurisdictions.