GPAI, FOUNDATION MODELS AND GENERATIVE AI IN EU AI ACT

The existence of AI systems that can be used to perform different tasks in different contexts is not new. The Commission's AI Act proposal rightly focused on applications of AI, rather than on the raw component models themselves.

Today a new generation of more capable and versatile AI systems has emerged, and the nomenclature has evolved accordingly - we now talk of "general purpose AI" (GPAI)[1] and "foundation models"[2], with "generative AI"[3] as a thematic subset. But all remain essentially multipurpose AI systems, and most will seldom, if ever, be used in high-risk settings.

Certain multipurpose models may need added precautions, and we welcome efforts to clarify how GPAI, foundation models and generative AI should be treated within the context of the AI Act. However it's vital to keep a sense of proportionality on any general restrictions and avoid being overly broad in scope or overly prescriptive in ways that could limit development of tools for societally beneficial applications. In practice this will require a clear focus on high-risk applications.

This note lays out some specific concerns and proposed recommendations in relation to the AI Act's treatment of GPAI, foundation models and generative AI.

1. The regulation of GPAI should focus only on the most capable foundation models when they are deployed for high-risk uses.
2. Requirements for Generative AI should be proportionate and apply to those best-placed to implement them.

1/ The regulation of GPAI should focus only on the most capable foundation models when they are deployed for high-risk uses

Council and Parliament have put forward different approaches for how GPAI and foundation models should be treated in the AI Act, but neither is ideal:

- The Council proposed that GPAI be regulated only if it could be deployed in high-risk AI applications, but left the concrete requirements to be clarified in a later implementing act.
    - *Why this is problematic*: The scope of the rules is drafted too broadly as it risks applying to all types of GPAI systems, many of which will have been on the market for some time and are well-understood to not present particular risks. Moreover, delaying the specification of GPAI requirements to future implementing legislation

[1] The Council and Parliament give different definitions for GPAI, although they are similar in spirit. The Parliament's definition is broadest: "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed". The Council's definition is more detailed: "an AI system that - irrespective of how it is placed on the market or put into service, including as open source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems".
[2] Only Parliament used this term, which they defined as "an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks"
[3] In their proposed revision to Article 28(b) Parliament specifies that generative AI is "foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video."

would extend the period of uncertainty for the GPAI ecosystem.

- In contrast, Parliament proposed to impose substantive high-level requirements only on foundation models regardless of specific application - on the theory that they represent the largest and most capable GPAI systems, and thus should be held to higher standards.
    - *Why this is problematic:* Although the suggested requirements may be sufficiently high level to provide leeway on implementation as technical standards evolve, it would risk overregulating foundation models that are only ever deployed in low-risk contexts (like one used to power an email spam filter) or sold to a third parties on the contractual condition that it cannot be deployed for high-risk applications. Alterations are also needed to the precise wording of certain requirements to make them fit-for-purpose.

How to fix:
The **scope of regulation for GPAI** should combine key elements of the Council and Parliament proposals, by **focusing on foundation models only when used in high-risk contexts.** This would be the most proportionate approach and aligned with the risk-based framework of the wider AI Act.

It is preferable to specify details of requirements directly in the AI Act (rather than in subsequent implementing legislation), to provide earlier clarity and because the appropriate scoping is so closely intertwined with the nature of requirements imposed. In order to future-proof the AI Act, and allow for broader development and adoption of foundation models in Europe, **regulatory requirements need to be sufficiently broad, flexible and adaptable**. The following aspects as proposed by the Parliament deserve cautious consideration:

- *Art 28b(2)(a)*: The **"rule of law" and "democracy" have differing interpretations** even between Member States. Governments, not the private sector, should define and evaluate risks in relation to these broad concepts. Therefore, **references to these terms should be removed** throughout the legislation. As a general observation, **risk mitigation efforts will be most meaningful at the application level** rather than at the model level. Techniques for reducing risk are not generic, and mitigations ideal for one application can be a hindrance to another with a different purpose and operational context. As an example, mechanisms to block abusive outputs would be helpful for a chatbot, but problematic for counter-abuse classifiers.

- *Art 28b(2)(b)*: Foundation models are trained on a wide variety of inputs drawn from the open web. This leads not only to more capable systems with more accurate outputs but also increases the versatility and representativeness of the models (including the limitation of bias). In some cases, it may be necessary to train on objectionable content precisely so that a model can learn to avoid outputting similar statements. In this context, **data governance should be understood as a conceptual process rather than a requirement of specific actions, e.g. vetting individual sources or webpages**. Suitable steps may include examining the possible biases associated with different types of data sources (e.g., online vs digital version of a national library).

- *Art 28b(2)(a) and (c):* A high level of expertise is required to effectively shape, oversee and assess various aspects of foundation model safety. Providers understand that robust red-teaming is essential for building successful products, ensuring public confidence in AI, and guarding against potential security and fundamental rights issues. As this is a rapidly developing field, we would caution against prescriptive requirements that risk limiting innovation and progress in this area. Providers, deployers and authorities should invest in research, drawing on domain experts on possible risks.

- *Art 28b(2)(d):* We support efforts to raise transparency about the energy and resource use of AI. Before contemplating any environmental requirements for AI, legislators should first **evaluate how any standards might be best designed and ensure consistency with existing EU environmental legislation** (e.g., the Energy Efficiency Directive in relation to data centres, the Ecodesign Regulation, or the Corporate Sustainability Reporting Directive). Sectoral product safety legislation, like the AI Act, is not the appropriate instrument.

- *Art 28b(2)(d):* While we agree with the need to ensure the cybersecurity of foundation models, **one-size-fits-all requirements concerning "appropriate levels" are generally ill-suited** to the regulation of AI systems with such a wide range of applications as foundation models. This is the case particularly for **performance and predictability** (where different metrics and benchmarks would apply to writing poetry versus legal contracts or chemistry formulae). In reality, it is the providers and deployers of the AI systems based on these foundation models who will need to take additional steps to align performance and predictability metrics to specific use cases and comply with relevant obligations if their AI products are high-risk.

2/ Requirements for Generative AI should be proportionate and be the responsibility of those in the best position to implement them

Generative AI is not "high risk" in and of itself — it would only become so if used in a specific context deemed high risk by the AI Act. The following considerations can help develop such applications in a way that increases consumer trust and transparency. The key is for such requirements to be scoped in a way that is proportionate and practical, and to avoid blurring lines of responsibility.

While all three institutions agree that generative AI is not inherently 'high risk', there are some differences in the details of what is being proposed. The Commission and the Council have put forward transparency requirements for Generative AI - specifically, for users to be informed when interacting with AI (unless it is obvious), and for artificially generated content to be labelled when it involves a 'deep fake'. The Parliament has gone further by proposing additional transparency requirements, additional data disclosures for training data, as well as additional safeguards to ensure that AI-generated content complies with Union law.

While many of the proposed requirements are sensible in spirit, changes are necessary to make them workable, aligned with other regulations (such as the Digital Services Act), and applicable to the appropriate actors in the value chain.

How to fix:
Any specific requirements for generative AI should be decoupled from more general rules for foundation models. Foundation models, by definition, are multipurpose and applied across a wide variety of applications. Providers typically have little visibility into (and control over) the context of their deployment in specific downstream applications, including various generative AI products.

The definition of generative AI should be tightened to more accurately reflect the types of AI the legislator is aiming to capture, namely systems that are intended to generate "complex content" for "direct consumption by natural persons". In terms of the specific obligations under consideration:

- Content safety requirements: Ensuring that EU content rules are respected is key, but Parliament's proposal to impose restrictions on foundation models[4] is problematic for several reasons. Different jurisdictions have different - sometimes even conflicting - content safety rules that could lead to contradictory requirements for providers. If this proposal were enacted, it would force developers to create a myriad of different foundation models to cater for different content safety regimes, slowing innovation, raising costs, and limiting deployment in Europe. A more workable approach would be to **apply any content safeguards at the application level where they can be implemented via fine-tuning and filtering techniques**. In addition, the requirement would benefit from a clearer scope of "content in breach of EU law"by identifying which specific EU Regulations and Directives providers should focus on and align with.

- Transparency/labelling requirements: More clarity is needed over who is responsible for helping users understand when AI may be contributing to content and in what circumstances (ranging from the accepted use of AI in applications like Google Search, Translate, and Maps, to potentially deceptive social media accounts). As a general point, **responsibility for transparency should lie with those in the best position to assess and most effectively mitigate risks**. More specifically:
    - The **mandatory labelling of AI-generated content should remain limited to the very specific category of deep fakes**, as envisaged in Art 52(3). And, of course, labelling should exclude any artistic, creative and similar works as envisaged in the Council's General Approach. Otherwise there is a real risk of 'labelling fatigue' (akin to cookie consent fatigue), where most online content ends up being labelled as "AI generated" and viewers stop paying attention to labels. In addition, legislators should not pre-empt sector-specific self-regulation (such as in advertising, healthcare, or financial services), where targeted approaches to labelling already exist or are under development.

[4] Art 28b(4)(b) proposes an obligation to "train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law".

- **Creators** of AI-generated content ultimately determine what they will publish and so should be responsible for applying appropriate deep-fake labels.
- Generative AI (or potentially foundation model) providers may be able to play a supporting role where relevant and technically feasible (e.g., through tags or watermarks for AI-generated content). It is currently impossible, however, for providers to ensure that such measures remain in place in the final format. Watermarks in imagery can be cropped or edited out, while watermarks in text, especially short text, can be even more challenging. As labelling techniques and their impact are still being explored, **we would urge that labelling by providers remain voluntary at this stage**.
- Independently, the requirement under Art 28b(4)(a) to **inform users they are interacting with a generative AI system** should be the responsibility of the **deployers** (and **not** the providers of the foundation model or generative AI). The deployers control the layout of a website or app and can integrate the messaging most appropriately for their audience and to suit local laws. For example, in the case of Bing AI it is the deployer (Microsoft) who is best placed to ensure that the end-user is properly informed (e.g., through website notices and warnings). The provider of the generative AI system or foundation model (OpenAI in this example) do not have the same (or perhaps any) level of access to the end-user.

- Copyright disclosure requirements: **The proposal to require publication of a "sufficiently detailed" summary of copyrighted training data is unnecessary and difficult to implement, and should be removed.**
  - There is no regulatory gap for copyright protection in relation to text and data mining (TDM) that warrants the imposition on new disclosure requirements in the AI Act. The EU's Copyright Directive already provides an option[5] for publishers to opt out of TDM. Since publishers can easily opt out of TDM *ex ante*, introducing an *ex post* transparency requirement is unnecessarily burdensome. There are also long-established IP enforcement mechanisms that publishers can use to obtain a court order to compel alleged infringers to disclose relevant information.
  - The concept is also out of date. Most foundation models have for years drawn training data from the constantly evolving and virtually unlimited open web, rather than discrete offline datasets. Preparing exhaustive and accurate (i.e., up-to-date) summaries of training data from the whole open web is hardly possible.
  - Finally, any summary of the *use* of training data in foundation models is valuable know-how that constitutes a trade secret. It is therefore commercially unreasonable and a violation of existing Member State IP protections to require public disclosure of such information without a court order tailored to a specific claim. Such disclosure could also be misused by malicious actors, or lead to leakage of technology that could be misused by malicious actors.

[5] Specifically, TDM is addressed in Article 4 of the EUCD, which was recently confirmed by the Commission to reflect an appropriate balance between rights holders and the facilitation of TDM. More generally, the EU system has a robust system of copyright protection and enforcement, in the shape of 12 copyright directives, 2 copyright regulations, and one Enforcement Directive which is applicable to all IP, including copyright.