September 2023

DOT Europe discussion paper on generative AI

Introduction

Discussions on the AI Act over the course of 2023 have been largely dominated by the development and increasing use of generative / foundational AI models, generating increased interest in the risks and opportunities of these technologies. At the same time, the rapid development of these technologies and the different stages of development and understanding during which the EU institutions adopted their respective negotiating positions have resulted in a varied and potentially contrasting approach to the regulation of technologies which are in constant development.

In the Parliament, MEPs have inserted an entire section in their final position with requirements specifically dedicated to generative AI, and many provisions not originally covered by the Commission's proposal. On the other hand, the Council, which adopted their negotiating mandate at an earlier stage of the discussion opted for an approach focused on "General purpose AI" which can be considered similar to foundation models. The significant level of attention placed on this issue, coupled with the quick pace of discussions – without the time required for a thorough legal and economic assessment of the provisions – has therefore resulted in some misconceptions around these technologies.

As the topic is likely to feature prominently in trilogue negotiations in Q4 2023, as well as at the international level, DOT Europe would like to take this opportunity to provide useful clarifications for policy makers to inform the ongoing discussion

Summary of recommendations:

- Clarify Article 28 to ensure provisions are workable in practice.
- Rules should reflect language adopted by the EP on the state of the art.
- Exclude obligation for foundation models on rule of law and democracy and energy.
- Support the Council's aim to focus on the highest risk uses when regulating GPAI/foundation models.
- Consider wording by the Council on the sharing of information in Article 4(b)(5).
- Providers should be able to explicitly exclude all high-risk uses in the instructions of use or information.
- Avoid combining the EP and Council Positions in terms of provisions on foundation models and GPAI.
- Predictability and interpretability: Responsibility of providers should solely extend to the development of the model, not to how it may be further affected once deployed.
- Remove the copyright-related obligations on foundation models (Article 28b).
- Remove the pre-notification provisions in Article 6.
- Retain the EP version of Article 52(3)1, with additional changes to remove "text" and specify that it would only cover artificially generated content disseminated to the public.
- Obligations to inform users under Article 28b(4)(a) should not exclusively be placed on one single actor.
- Deep fake content: Policy makers should maintain the wording proposed by the EP on article 52.3 and 52.3a.
- The definition of deepfakes should be kept sufficiently targeted.

What is meant by "generative AI," "foundation models," "large Language Models" and "General Purpose AI"?

1. **Generative AI** is a type of machine learning model that can generate new content types, such as text, images, music etc. based on user inputs. The underlying AI model has been trained through the observation and recognition of patterns from a vast array of data. The technology then uses this training to predict what would be a convincing answer to a prompt. As such, it does not function as an information system. The information it presents is not held in a database, which means that these systems are not intended to function as sources of reliable information. Their primary use is to generate plausible output that appears convincing. As they are simply prediction engines, they may also create different outputs in response to the same prompt by the user.
2. **Large language models (LLMs)** are a type of AI system which is trained on text data and has the ability of generating language responses to inputs or prompts. LLMs are often considered synonymous with foundation models. A key distinction can be drawn where the LLMs often refer to language-based systems whilst a foundation model is broader.
3. Generative AI and LLMs are both a type of **foundation model** which is the underlying model used to train these systems. They form the basis (foundation) of the systems hence their name.
4. **General Purpose AI Systems** (GPAI) are systems that can be used in and adapted to a wide range of applications, including some for which it was not intentionally and specifically trained. This term is often used interchangeably with foundation model, however, not all GPAIs can be characterised as foundation models and vice-versa as there exist certain nuances between the two. For example, GPAIs may use different approaches in their training than those generally used by foundation models focusing on pure transfer learning for example.

Current state of policy debate

In broaching the topic of generative AI and foundation models at the EU level, it is important that such discussions do not take place and evolve in a vacuum and consider the current wider international context of the debate. In this context, it is important to note and take stock of recent milestones in this debate, namely, the ongoing voluntary industry efforts from leading AI companies, or the future EU-US-driven AI Code of Conduct expected to be adopted by the G7.

International efforts on AI

At the multilateral level, there are several ongoing efforts seeking to align AI governance and measures globally. At the G7 in Japan, political leaders set up the "Hiroshima AI process" tasked with discussing AI governance, IP rights protections, transparency and action plans to address harms such as misinformation. This effort will most likely result in what has been a US-EU-driven international effort on a voluntary AI Code of Conduct focusing on international standards on risk audits, transparency and other requirements for companies developing AI systems and which would then be put before G7 leaders as a joint proposal.

Ongoing Industry Efforts

Industry has also stepped up with a variety of voluntary efforts seeking to encourage safe, secure, and transparent development of AI technology. This has been the case collegially with cross-industry or public-private commitments (i.e., the recently launched Frontier Model Forum – supported by, among others, Microsoft and Google - or the voluntary commitments agreed upon by seven leading AI companies – including Amazon - with the White House on AI). It has also been the case at the individual company level with leading companies within this realm continuing and strengthening their ongoing work seeking to ensure a safe and responsible innovative development and deployment of AI (i.e. see Meta's Responsible use Guide for developers).

Comments on the trilogue negotiations

1. Obligations on foundation models

Unworkable and disproportionate obligations: The European Parliament has imposed a **wide range of obligations** on foundation models that may be unmanageable for developers and deployers of foundation models, including European developers aiming to invest and innovate in this flourishing market. Such obligations often lack clarity and may be overly broad which would result in a disproportionate burden for developers. These include data governance requirements that may contrast with the principle of data minimisation; requirements for performance levels and energy use; and obligations to identify, reduce and mitigate risks to democracy, the rule of law, or the environment. Moreover, it is unclear whether such obligations may, in practice, provide the needed assurances of improved confidence. Furthermore, it is important that policymakers take the time to consider the obligations imposed to ensure that the burden is shared between all actors of the AI supply chain and assigned to the actor with the necessary knowledge needed to implement those obligations.

Overlap with copyright law: Moreover, MEPs suggest imposing additional obligations on foundation models, in particular on the **use of training data protected under copyright law** and associated disclosure obligations. These provisions present significant overlap with Article 4 of the Copyright Directive, which allows rightsholders to opt-out of text and data mining. This makes the obligations in the AI Act redundant. Furthermore, data collection from the web is done dynamically, meaning it is updated over time. This would render disclosure obligations impossible to meet due to the sheer amount of content that would have to be disclosed. Finally, "use of training" data is valuable know-how and may constitute a trade secret. At the same time, any public disclosure of the use could be misused by malicious actors without necessary IP and trade secret protections.

Sharing of information: The European Parliament has proposed amendments regarding the provision of information by developers of foundation models. DOT Europe supports the goal of Article 28b aiming to increase the sharing of information between stakeholders in the supply chain also with a view to ensuring all parties have the required knowledge to comply with relevant obligations under the Act. However, for such provisions to have an impact they would need to be workable in practice and thus would benefit from further clarification to avoid any potential legal uncertainty. For instance, it is not clear to whom the information needs to be disclosed, what the "specialisation of the foundation model" means and what a "sufficiently detailed summary of the use of training data'' might look like in practice. In this context, deployers of such systems should be provided by developers with the information needed in order to comply with their obligations under the AI Act. The Council provides for helpful wording in this regard in Article 4(b)(5).

Furthermore, it is reasonable for developers of such systems to forbid high risk uses or limit liability for actions by third parties. The Council provides for helpful wording in this regard in Article 4(c). DOT Europe would like to caution policymakers, however, on adopting a combination of the European Parliament and Council Positions in terms of provisions on foundation models and GPAI respectively. Doing so could render the helpful wording and approach by the Council outlined above overly burdensome. For example, a combination of the positions might result in much broader obligations on deployers for outcomes they would not be able to actively control. Finally, policymakers should also further consider a shared burden when it comes to requiring providers of foundation models to ensure appropriate levels of **predictability and interpretability** throughout a foundation model's lifecycle (Article 28b(2c)), as these factors are heavily influenced by decisions taken by application developers and therefore may be out of the model provider's control in certain instances.

DOT Europe recommendations:

- The risk-based nature of the framework should be safeguarded throughout trilogue negotiations.
- Article 28 should be further clarified to ensure that provisions are workable in practice and do not result in legal uncertainty.
- Obligations on transparency and data governance imposed on foundation models should be risk-based, proportionate to a foreseeable risk level and role of developers of foundation models as well as be attainable in practice, as in the Council text.
- These rules should also reflect language adopted by the European Parliament on the state of the art.
- The obligation for foundation models related to rule of law and democracy and energy use should not be included in the AI Act.
- As a general rule, any such additional requirements not covered in the original proposal by the European Commission should be subject to a thorough impact assessment.
- Overall, we encourage policy makers to reflect on the complexity of the supply chain. The chain of responsibility and ensuring this is assigned in the right place will ensure a balance and that use of AI does not become prohibitively risky.
- DOT Europe supports the Council's aim to focus on the highest risk uses when regulating GPAI/foundation models.
- Policymakers should consider helpful wording by the Council on the sharing of information outlined in Article 4(b)(5).
- Providers should be able to explicitly exclude all high-risk uses in the instructions of use or information. The Council provides helpful wording to this effect in Article 4(c)(1) of the Council text.
- In order for the Council wording outlined above to be effective and to avoid any potential contradictions, DOT Europe particularly urges co-legislators to avoid a combination of the European Parliament and Council Positions in terms of provisions on foundation models and GPAI respectively
- A compromise should be sought when it comes to ensuring appropriate levels of **predictability and interpretability**. The responsibility of providers should solely extend to the development of the model, not to how it may be further affected once deployed. i.e., if it is deployed in a way that was not foreseen or intended by the developer which have changed the way the model works or behaves for example.
- Policy makers should remove the copyright-related obligations placed by the European Parliament on foundation models from the AI Act (Article 28b).
- The pre-notification provisions in Article 6 (namely Article **6.2a** as introduced by the European Parliament) should also be removed, as these would create significant implementation challenges for operators and authorities alike.

2. *Obligations on the output of generative AI (article 52)*

Dynamic nature of text-based AI content: DOT Europe believes there is some benefit in requiring AI generated images and audiovisual content to be labelled in key scenarios, so that the public "knows the content" it is being exposed to. Reasonable measures to deter the misuse of new technology to deceive or defraud the public will benefit the health of democracy and the future of civic discourse. However, the addition of text-based content as added by the Parliament in Article 52(3) does not appear to be appropriate in the context of the requirement. The highly dynamic interaction between text-based AI generated content, as well as user amendment and refinement of those texts makes labelling such content in a meaningful way much more difficult than in cases involving similar technologies for the labelling of image/video-based content. It might be more appropriate to include text-based content under the provisions on informing consumers that they are interacting with an AI system (article 52.1). The mandatory labelling of AI-generated content should remain limited to the very specific category of deepfakes, as envisaged in Art 52(3). 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 consider sector-specific self-regulation (such as in advertising, healthcare, or financial services), where targeted approaches to labelling already exist or are under development.

Independently, the requirement under Art 28b(4)(a) to inform users they are interacting with a generative AI system would benefit from further distinctions. The obligation might not be best placed exclusively on the developer nor the deployer. From a technical point of view, exclusively introducing labelling requirements at the model level might not be appropriate. Policymakers should consider specific deployment scenarios. In the case of third-party deployers given multiple scenarios of use and practicalities of implementation the deployer might be best placed to ensure that the end-user is properly informed (e.g., through website notices and warnings) as the provider may not have the same level of access to the third-party end-user. This is undoubtedly different when it comes to scenarios in which the developer also deploys the system. Further consideration should also be given to different needs in terms of technical implementation. The labelling of text for example is less burdensome and/or difficult to implement at the level of use than other types of content. As further outlined below, this should be read in conjunction with our recommendation that only generative AI systems which interact with end-users should comply with Article 52.1.

Deepfakes labelling: We maintain that a more nuanced approach would be needed on the issue of deepfakes. We agree that transparency should be the priority when it comes to deepfakes, and we particularly support the aim of Article 52 to give more clarity to users. However, while we support labels that alert users of deepfakes, it is important that these labels are not prescriptive as to obstruct a user experience on a given service. We also caution against an overly broad definition and interpretation of the definition of deepfakes to avoid capturing unintended technologies and services that do not present the same risk and which focus on the creative and artistic freedom of end-users. This is, for example, the case of AI-powered filters which have become popular and sought-after among end-users and which use machine learning techniques to design augmented reality filters. This is also the case for distinct practices within the different realms of the creative sector which provide for vastly different risk levels and which may not possess the characteristics of a deepfake. A clear example of this is the use of computer-generated imagery (CGI) used in the film or video game sector for example. The definition of deepfakes and the priorities in enforcement by regulators should therefore reflect and consider these practices and recognise the varied risk levels they pose.

Distinction between general public / enterprise use: The transparency requirement in Article 52(3)1 would benefit from taking into account the distinction between generative AI services disseminated to the public (i.e., as defined in the Digital Services Act) and those provided in a (closed) enterprise environment or in the context of (business or private use of) productivity software.

DOT Europe recommendations:

- DOT Europe calls on the co-legislators to retain the Parliament's version of Article 52(3)1 in the final text, with additional changes to remove "text" and specify that the obligation would only cover artificially generated content disseminated to the public.
- Obligations to inform users under Article. 28b(4)(a) should not exclusively be placed on one single actor. Differing deployment scenarios should be considered as well as technical implementation with a view to a more collaborative supply chain based approach.
- Regarding "deepfake" content, policy makers should maintain the wording proposed by the European Parliament on Article 52.3 and 52.3a, which we believe adequately supports artistic freedom and supports the necessary distinction between technologies, services and risks outlined above.
- The definition of deepfakes should be kept sufficiently targeted to avoid capturing unintended technologies and services that do not present the same risk and which focus on the creative and artistic freedom of end-users.

Conclusion

Trilogue discussions on the AI Act are expected to cover generative AI systems in the second half of 2023. The political imperative to achieve an agreement as soon as possible is very strong and policy makers will likely try to work through the text quickly. DOT Europe believes that speed should not be to the detriment of quality and urges trilogue participants to adequately consider the various aspects and nuances of generative AI to ensure a balanced, workable text.