July 2023

FLI Position Paper: AI Act Trilogue

Mark Brakel, Risto Uuk
policy@futureoflife.org

Eminent pioneers in the field of AI research and development[1] **as well as contemporary innovators**[2] **are increasingly issuing warnings over the safety threat from advanced AI systems.** These warnings further underline the need for a strong AI Act. We believe that there is a pressing need for this legislation so that AI development is steered away from ongoing and growing harms.

As the debate on the AI Act reaches its final stage, **we should remember that it is** investment in computational power, data and talent, **not regulation, that will determine the fate of AI development in Europe**. For example, non-EU capital London hosts more AI talent than the combined share of Paris, Berlin, Madrid and Amsterdam.[3] Similarly, one Californian AI corporation (OpenAI) has 25 times[4] (!) the advanced computing capacity of the entire United Kingdom combined, which houses Google DeepMind and is thus likely ahead of any single EU Member State.

If Europe wants to catch up, regulators should focus on investment and talent, not on undermining basic legal safeguards. In fact, solid regulation can help European developers carve out a global brand for trustworthy AI. The commercial aviation sector provides an example of this. Heavy regulation in this sector has encouraged a 'race to the top' in safety, leading to an 83% decrease in fatality risk between 1998 - 2008 alongside a 5% annual increase in passenger kilometers flown. Our joint analysis with the Boston Consulting Group further sets out this economic case for AI safety.[5]

The proposals of the Parliament and the Council provide a solid foundation for an AI Act that encourages this 'race to the top' for advanced AI systems. The Parliament's proposal, in particular, clearly allocates responsibilities along the AI value chain and gives European SMEs the insights they need about the systems that they incorporate in their final products.

The Parliament proposal rightfully assigns most responsibility to the builders of general-purpose AI systems, because they have the necessary financial resources and knowledge to comply. A clearly delineated AI value chain also makes it easier for downstream deployers to switch to emerging European providers when these have matured and mitigates the risks of AI value chain lock-in (as we see in European cloud computing today).

Some tweaks are however necessary such that the Act generalises from today's focus on ChatGPT; the law should also capture systems that can act (for example by creating dangerous chemicals) or that engage in planning over a longer time horizon. In tweaking the law, FLI encourages the co-legislators to maintain a sufficiently broad definition of general-purpose AI systems that fully captures systems that threaten our information environment, elections, or global security.

This document outlines FLI's updated position on the AI Act taking into account the positions of the co-legislators.

1 New York Times, 'The Godfather of A.I. Leaves Google and Warns of Danger Ahead.' https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

2 The Guardian, 'We are a little bit scared: OpenAI CEO warns of risks of artificial intelligence.' https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4

3 Sifted, 'Where to find Europe's best AI engineers.' https://sifted.eu/articles/europe-engineering-talent

4 James W. Phillips, 'Securing Liberal Democratic Control of AGI through UK Leadership.' https://jameswphillips.substack.com/p/securing-liberal-democratic-control

5 Gupta et al., 'Emerging AI Governance is an Opportunity for Business Leaders to Accelerate Innovation and Profitability.' https://techpolicy.press/emerging-ai-governance-is-an-opportunity-for-business-leaders-to-accelerate-innovation-and-profitability/

Contents

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

Recommendations

- **General purpose AI systems include foundation models and generative AI systems**: the definition and regulatory treatment should reflect this to provide legal clarity.
- Providers of general purpose AI systems (which include foundation models and generative AI) should undertake "know-your-customer" **(KYC) checks** throughout the entire product's lifetime to ensure that widely dispersed harms from these systems are mitigated at source from those that understand them best.
- Providers of general purpose AI systems should undergo **third-party conformity assessments** as envisaged for "real-time" remote biometric identification systems. These frontier systems present emerging capabiilities with numerous identified risks and unpredictable risks should be managed ex-ante. Trusting private corporations to self-police will not sufficiently protect health, safety, and fundamental rights in the EU.
- **Providers of general purpose AI systems should not be allowed to evade their responsibilities**, as the largest and most well-resourced companies in the world, with loopholes that allow them to avoid regulation by declaring that their systems should not be deployed in a high-risk use case.
- As the most advanced technology in society, **AI demands a new EU agency** to ensure effective coordination of enforcement among Member States. This AI Office should consult with all relevant stakeholders, including civil society, in dialogues with GPAI providers about planned releases of increasingly sophisticated AI systems and when reviewing the legal regime governing GPAI.

General Purpose AI: Foundation Models and Generative AI

Definition

SUGGESTION

Article 3 (Definitions)

'General purpose AI system' means an AI system that – irrespective of how it is placed on the market or put into service, including as open-source software – can be used in, and adapted to, a wide range of distinct and downstream tasks, including some for which it was not intentionally and specifically designed. General purpose AI systems can be foundation models to other narrower single purpose AI systems and can also generate new and original content such as images, videos, text, and audio.

Corresponding recital (new)

General purpose AI systems include both unimodal and multimodal systems that can be trained through different methods, and currently encompass many fields, such as natural language processing, computer vision, speech, and robotics, among others. General purpose AI systems can also include foundation models, because they can be both used in standalone systems and as the base infrastructure upon which many other single-purpose AI systems can be built and adapted (e.g. fine-tuned) to a wide range of distinct downstream tasks. General purpose AI systems can also include some generative AI models that can create new and original content, such as images, videos, text, and audio. Given their foundational and multifunctional nature, general purpose AI systems with emergent capabilities and unpredictable risks can pose significant systemic harms to society as a whole, including democracy and the rule of law, which should be adequately addressed in risk mitigation measures and the governance of training datasets.

JUSTIFICATION

This combined definition of the Council and European Parliament texts clarifies that foundation models and generative AI systems fall under the category of general purpose AI systems (GPAI). GPAI systems have many more intended – and unintended – uses than single purpose AI systems. Their defining feature is performing numerous distinct tasks, with distinct being crucial to capturing only the most disruptive systems. **Foundation models**, such as OpenAI's GPT models, **are a form of GPAI, serving as the basis for differently purposed downstream user-facing applications** (ChatGPT, Bing, etc.).

While all GPAI systems currently possess generative capabilities, among others, not all generative AI systems are considered general purpose. Generative AI can be narrowly designed for that purpose, or be a capability of GPAI. And GPAI systems can go beyond generative capabilities: GPT-4, for example, is not merely generative. The system can analyse legal texts, produce code in a variety of languages[6], and control robotic functions.[7]

The Act should also include systems that will be sold on the internal market soon. Researchers have recently used language models not just to generate text and images ('generative AI') but also to carry out actions. For example, a number of chemists recently fine-tuned a large language model so that it could synthesise molecules.[8] Similarly, the CEO of Google DeepMind - the main

6 OpenAI, 'Introducing ChatGPT.' https://openai.com/blog/chatgpt

7 Microsoft, 'Autonomous Systems and Robotics Group.' https://www.microsoft.com/en-us/research/group/autonomous-systems-group-robotics/articles/chatgpt-for-robotics/

8 Andres Bran et al., 'ChemCrow: Augmenting large-language models with chemistry tools.' https://arxiv.org/abs/2304.05376

rival to current market leader OpenAI - expects its new 'Gemini' system to be able to execute a wide range of additional tasks including "planning" and the "ability to solve problems".[9]

FOUNDATION MODELS

If co-legislators maintain the European Parliament's inclusion of a definition on foundation models, this may be best defined as follows: 'foundation models are AI models designed for generality of output, and can accomplish or be adapted to accomplish a wide range of distinct (downstream) tasks.' Unlike the definition provided by the Parliament, this definition does not rely on the term "broad data at scale", which may prove problematic when providers train their models in simulated environments that rely less heavily on large scale training datasets to perform general functions.

Distinguishing between foundation models and other GPAI systems without downstream applications is a risky approach, however, because it could create distortionary incentives. If the AI Act creates two different classes of GPAI (foundation and non-foundation) with different regulatory burdens, GPAI providers may temporarily bar European SMEs from incorporating their systems in an attempt to avoid extra requirements. Many AI developers don't decide beforehand whether their system will be put on the market directly or through downstream applications, and the Parliament's approach risks incentivising one business model over another. FLI therefore supports the Council approach, which does not distinguish between different classes of GPAI.

GPAI SYSTEM	FOUNDATION MODEL?	ONLY GENERATIVE
Claude (Anthropic)	No, only provided as stand alone on market directly	Yes
GPT-4 (OpenAI)	Yes	No, can be used as an agent[10]
Gemini (DeepMind, not yet released)	?	No

Figure 1: Table with classification of some leading GPAI systems

9 Wired, 'Google DeepMind's CEO Says Its Next Algorithm Will Eclipse ChatGPT.' https://www.wired.com/story/google-deepmind-demis-hassabis-chatgpt/

10 Wikipedia, 'Auto-GPT.' https://en.wikipedia.org/wiki/Auto-GPT

Value Chain

SUGGESTION

FLI supports proposals by both co-legislators to ensure that any natural or legal person shall be considered a provider if they place on the market or put into service a general purpose AI system as a high-risk AI system or as a component of a high-risk AI system (Article 23a in the Council text), or they make a substantial modification to a general purpose AI system in a manner that changes it from non-high-risk to high-risk (Article 28 in the Parliament text). **These obligations should apply to all general purpose AI systems, which include foundation models and generative AI systems.**

SUGGESTION

FLI supports the European Parliament's efforts to regulate providers of foundation models and generative AI by providing robust governance, reporting and transparency obligations detailed in Article 28b. However, these **obligations should apply to all general purpose AI systems**, which include foundation models and generative AI systems.

We also recommend adding an additional clause 5 specifying that GPAI providers need to conduct Know-Your-Customer (KYC) checks to be able to enforce their terms and conditions:

28b.5. Providers of such systems shall undertake 'know your customer' checks throughout the system's life cycle. This should include capturing downstream deployers' intended uses, planned modifications to the systems, and undertaking regular checks thereafter to identify if the system is being used as declared. Once these checks identify a risk, the provider shall ask a downstream deployer to change the way they use the system, restrict or withdraw access, or take any other appropriate action to mitigate the identified risk.

Corresponding recital

"Know-Your Customer" (KYC) checks should be limited to what is necessary and proportionate to ensure that downstream deployers are using GPAI models according to the instructions of use, including changes that may amount to a substantial modification, which would be more easily determined by the GPAI provider. They should respect the legitimate interests of both parties, in particular in relation to the protection of trade secrets and confidential information. KYC checks should reduce the burden of ongoing risk monitoring for downstream deployers, the overwhelming majority of which will be SMEs. KYC checks should not allow GPAI system providers to gain access to information from downstream deployers that could lead to competition concerns between the parties.

SUGGESTION

Another means of achieving effective ongoing risk mitigation, including situations that may lead to a substantial modification, is to extend Article 12 "record-keeping" requirements to GPAI system providers to ensure that they develop GPAI systems that enable the automatic recording of events ('logs') throughout the lifecycle. In line with Article 29 deployer obligations on implementing human oversight, downstream deployers would then be required to inform the GPAI provider of events presenting a risk while suspending use of the GPAI system. This would allow deployers to flag to GPAI providers any emerging significant risks for society as a whole, which usually emerge gradually rather than immediately, and which may only be discovered in ongoing post-market monitoring.

JUSTIFICATION

While GPAI systems have vast potential for improvements in many areas of life, the scope for harm is also substantial. This technology already exhibits a tendency toward amplifying

entrenched discrimination and biases, further marginalising disadvantaged communities and diverse viewpoints[11] to the detriment of democratic discourse, as recognised by the European Parliamentary Research Service (EPRS).[12] **The same AI systems could also threaten national security, for example by facilitating the inexpensive development of chemical, biological, and cyber weapons by non-state groups.**

Despite continual updates to improve accuracy, ChatGPT has a tendency to generate fabricated answers,[13] particularly in longer outputs, in ways that the developers were unable to predict, [14] with no built-in mechanism to signal this to the user.[15] Malicious actors can exploit this tool to spread disinformation and discord, especially among users unfamiliar with the subject matter.[16] The automation and scalability of such propaganda pose potentially destabilising effects on Europe's democratic institutions.[17]

The European Parliament's tailored obligations for foundation models are promising and pertinent. To ensure forthcoming models are not missed by narrowly scoping only foundation models, **these obligations should cover all general purpose AI systems, including foundation models and generative AI.**

To future-proof these provisions, GPAI providers should be required to conduct "know-your-customer" (KYC) checks, ensuring downstream deployers use their models according to the instructions of use. Since downstream applications will be innumerable, risks can stem from multiple sources, similar to financial markets in which KYC checks are a common best practice. Yet large GPAI providers could implement checks without significant burdens, as they would likely only apply to a select number of customers.[18] While OpenAI's usage policies highlight prohibited uses, SMEs may still unknowingly generate risks when using GPT.[19] In fact, Microsoft has already proposed a version of KYC checks applied to sensitive areas,[20] while OpenAI are also in favour of such measures.[21] Equally, if GPAI providers built-in mechanisms that would allow for record-keeping, as is the case for high-risk AI, then deployers could report failures or risks, similar to submitting software bug reports.[22] Thus, it is both downstream monitoring and upstream reporting, which would mitigate risk in a holistic manner.

It is imperative that European policymakers seize this moment to regulate GPAI in primary legislation, rather than deferring it to implementing acts, which would further delay the implementation of such rules beyond the two-year transposition deadline. With the arrival

11 Abubakar Abid et al., 'Large language models associate Muslims with violence.' https://www.nature.com/articles/s42256-021-00359-2

12 EPRS, 'General purpose artificial intelligence.' https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf

13 OpenAI, 'Introducing ChatGPT.' https://openai.com/blog/chatgpt

14 Cybernews, 'ChatGPT's answers could be nothing but a hallucination.' https://cybernews.com/tech/chatgpts-bard-ai-answers-hallucination/

15 Search Engine Journal, 'OpenAI's ChatGPT Update Brings Improved Accuracy.' https://www.searchenginejournal.com/openai-chatgpt-update/476116/

16 Axios, 'Chatbots trigger next misinformation nightmare.' https://www.axios.com/2023/02/21/chatbots-misinformation-nightmare-chatgpt-ai

17 Eurasia Group, 'Top Risks 2023.' https://www.eurasiagroup.net/files/upload/EurasiaGroup_TopRisks2023.pdf

18 CLTR, 'Response to the UK's Future of Compute Review: A missed opportunity to lead in compute governance.' https://www.longtermresilience.org/post/response-to-the-uk-s-future-of-compute-review-a-missed-opportunity-to-lead-in-compute-governance

19 OpenAI, 'Usage Policies.' https://openai.com/policies/usage-policies

20 Microsoft, 'Governing AI: A Blueprint for the Future.' https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

21 OpenAI, 'Safety best practices.' https://platform.openai.com/docs/guides/safety-best-practices

22 Ada Lovelace Institute, 'Expert explainer: Allocating accountability in AI supply chains.' https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

of ChatGPT on the world stage as the fastest-growing consumer application in history,[23] the last six months underscore the urgency for regulation to keep up with this rapidly evolving technology.

Third-Party Conformity Assessments

SUGGESTION

FLI supports the Council's efforts to subject providers of general purpose AI systems to conformity assessment procedures. However, **GPAI conformity assessments should be based on an assessment of the quality management system and technical documentation, with the involvement of a notified body, referred to in Annex VII**, not on internal control, as currently envisioned (Article 43). Such conformity assessments should take due consideration of systemic risks to society as a whole, including democracy and the rule of law.

JUSTIFICATION

GPAI systems trained on massive datasets have unexpected (and often unknown) emergent capabilities.[24] As a single failure mode, flaws in GPAI systems can have far-reaching consequences across multiple sectors.[25] Cyber risks, such as data poisoning, where malicious actors feed the user interface with triggering data, can cause the model to produce further harmful outputs to many other users across society.[26] Humans may intentionally misuse these capabilities or, due to technical alignment failures, the models could inadvertently produce negative outcomes.[27] **Pre- and post-market obligations are essential due to the deliberate or accidental repurposing of these models for ends with varying and unpredictable levels of risk.**

Major general purpose AI developers, such as Anthropic[28] and OpenAI[29], acknowledge the need for independent audits prior to product launch. They recognise that these models are growing increasingly powerful and serve as a crucial step towards achieving AI that equals or surpasses human intelligence. **Independent auditing, combined with the attractiveness of the world's largest affluent consumer base, incentivises non-EU companies to comply with Union values.**

Currently, European companies have been unable to rival their American and Chinese counterparts in developing competitive general-purpose AI systems due to the extensive financial resources, concentration of human talent, and computational power required. Consequently, innovation has been limited to major corporations, resulting in market concentration around data and model ownership.[30] Given the head start of big incumbents, European companies are expected to maintain their reliance on large scale systems developed elsewhere, and inequalities in resource investment will likely lead to growing dependencies.[31] Considering these asymmetries, **the original developers** (almost exclusively large multinational companies

23 Reuters, 'ChatGPT sets record for fastest-growing user base.' https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
24 Samuel R. Bowman, 'Eight Things to Know about Large Language Models.' https://arxiv.org/pdf/2304.00612.pdf
25 Standford University, 'Reflections on Foundation Models. https://hai.stanford.edu/news/reflections-foundation-models
26 Rishi Bommasani et al., 'On the Opportunities and Risks of Foundation Models.' https://arxiv.org/pdf/2108.07258.pdf
27 Toby Shevlane et al., 'Model evaluation for extreme risks.' https://arxiv.org/pdf/2305.15324.pdf
28 Anthropic, 'Core Views on AI Safety: When, Why, What, and How.' https://www.anthropic.com/index/core-views-on-ai-safety
29 OpenAI, 'Planning for AGI and beyond.' https://openai.com/blog/planning-for-agi-and-beyond
30 Rishi Bommasani et al., 'On the Opportunities and Risks of Foundation Models.' https://arxiv.org/pdf/2108.07258.pdf
31 Ganguli et al., 'Predictability and Surprise in Large Generative Models.' https://arxiv.org/abs/2202.07785

headquartered in the US or China) **should undergo third-party conformity assessments to give more certainty to European SMEs incorporating these systems.** As the most well-funded companies globally, the original developeres are capable of absorbing the compliance costs needed to ensure trustworthy AI in the EU.

Furthermore, developers may deploy models without full comprehension of their potential harm, and downstream fine-tuning models on new data adds complexity to predicting and controlling their behaviours.[32] **We thus recommend third-party auditing of GPAI systems across a range of benchmarks for the assessment of risks,[33] including possible weaponization,[34] unethical behaviours,[35] dangerous capabilities[36], and systemic risks to society as a whole[37].** Accredited third-party auditors should sign off on mandatory certification before these high-risk systems can be deployed. Certification should only be granted if developers can demonstrate adequate risk mitigation measures, disclose tolerable residual risks, and comply with established protocols for minimising harm.

Loopholes

SUGGESTION

The Council text includes a provision that exempts providers of general purpose AI systems from their obligations provided they have "explicitly excluded all high-risk uses in the instructions of use" if they do not have "sufficient reasons to consider that the system may be misused" (Article 4c). **FLI encourages co-legislators to remove this clause because in practice it will allow large and well-resourced companies to evade their responsibilities**, while leaving less well-resourced SMEs downstream with all the compliance burden.

JUSTIFICATION

GPAI systems should be regulated regardless of whether they are later used in a high-risk use case. By allowing providers to divest their responsibilities simply by excluding high-risk uses in the instructions for use, even if they are fully aware that their systems present a significant risk, creates an unacceptable loophole that is clearly open to abuse. It shifts all requirements downstream, facing deployers with technically unfeasible obligations, as they did not make the original data and design choices and will face significant barriers to seeking redress if the GPAI system causes harm. **Big Tech providers are thus best placed to assess risk in their own GPAI systems.**

Deployers are best placed to meet requirements for specific high-risk use cases, including human oversight, use-case-specific quality management, technical documentation, logging, and any additional robustness and accuracy testing. These obligations are particularly important for novel use cases that providers may not foresee.

32 Ganguli et al., 'Predictability and Surprise in Large Generative Models.' https://arxiv.org/abs/2202.07785

33 Rishi Bommassani, Percy Liang, & Tony Lee, 'Language Models are Changing AI: The Need for Holistic Evaluation.' https://crfm.stanford.edu/2022/11/17/helm.html

34 OpenAI described weaponization risks of GPT-4 on p.12 of the "GPT-4 System Card". https://cdn.openai.com/papers/gpt-4-system-card.pdf

35 Alexander Pan, et al., 'Do the Rewards Justify the Means? Measuring Trade-offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark.' https://arxiv.org/abs/2304.03279

36 Toby Shevlane et al., 'Model evaluation for extreme risks.' https://arxiv.org/pdf/2305.15324.pdf

37 Andrew Critch & Stuart Russell, 'TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI.' https://arxiv.org/abs/2306.06924

SUGGESTION IN EUROPEAN PARLIAMENT TEXT	
Article 3 (*Definitions*) 'significant risk' means a risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its ability to affect an individual, a plurality of persons or to affect a particular group of persons, or society as a whole.	Article 6 (*Classification rules for high-risk AI systems*) 2. In addition to the high-risk AI systems referred to in paragraph 1, AI systems falling under one or more of the critical areas and use cases referred to in Annex III shall be considered high-risk if they pose a significant risk of harm to the health, safety or fundamental rights of natural persons. Where an AI system falls under Annex III point 2, it shall be considered high-risk if it poses a significant risk of harm to the environment, or to society as a whole.

JUSTIFICATION

Some AI systems have arguably caused greater harm at the aggregate, rather than individual, level, such as the Cambridge Analytica micro-targeting of voters in the Brexit referendum. **FLI therefore recommends adding "society as a whole" to the definition of significant risk (Article 3) and high-risk AI classification rules (Article 6)** to cover systemic risks to both individuals and society, including democracy and the rule of law, as in the recently adopted Digital Services Act.

Governance

AI Office

SUGGESTION

FLI **supports the European Parliament's proposal for an AI Office**. It was encouraging to follow the Council's addition to the AI Board which foresees a standing subgroup that would serve as a stakeholder platform (Article 56.3.2 in the Council general approach). However, it is important that the EU establishes an independent body that not only consults regularly with industry and civil society, but also institutionalises dialogues between regulators and developers, and issues annual reports on the development of foundation models, including policy options specific to them, as foreseen by the European Parliament (Article 56b).

While this would provide for effective oversight and coordination at the EU level, it could be reinforced by **requiring developers of all general purpose AI systems**, including foundation models and generative AI, to not only comply with the above, but also to provide information on release plans for systems currently under development.

Moreover, the AI Office should include civil society in institutionalised dialogues and provide a mechanism for vetted researcher access to advanced AI models.

JUSTIFICATION

By creating an independent EU AI body or agency, EU policymakers will ensure that the AI Act does not become a "paper tiger". This entity should ensure that the complicated process of enforcement does not lead to fragmentation of the single market and varied protection levels across the Union for citizens.

Strong centralised enforcement is essential if co-legislators retain the European Parliament's provision for self-regulation, which allows providers to release high-risk AI systems based on their own judgment of significant risk, while awaiting a response to their reasoned notification (Article 6.2a and 6.2b). This is a potential conflict of interest whereby the entity profiting from a product decides if it is safe to sell.

While EU data protection authorities have considerable expertise, AI presents risks beyond privacy. An AI-specific authority, such as the Office, is appropriate to manage the more diverse risks tied to ubiquitous AI adoption across sectors. Institutionalising dialogues within the AI Office between regulators, GPAI providers, and civil society is a necessary part of addressing risks from larger, perhaps more unpredictable models scheduled for release.

Moreover, the appropriate governance regime for the AI Office should also learn lessons from social media. Years after they were first deployed, we still do not understand the impact on society of many basic social media algorithms. The Digital Services Act rightfully provides public interest researchers with access to data held by major tech firms, and **the AI Office should be similarly empowered to provide vetted researchers with access to black box AI systems.**[38]

38 John Albert, 'A guide to the EU's new rules for researcher access to platform data.' https://algorithmwatch.org/en/dsa-data-access-explained/

SUGGESTION

FLI supports the European Parliament's additions to Article 84 on evaluating the Regulation's effectiveness, especially in light of future developments. Namely, that:

- the Commission should consult the AI Office and relevant stakeholders when producing an evaluation report every two years, which should be made public;
- these reports should review the legal regime governing foundation models;
- the Commission should assess if the AI Office is sufficiently resourced; and
- account for democracy and the rule of law when considering amendments to the Regulation.

These evaluations could be strengthened by including a review of standards developed for general purpose AI by the Commission, AI Office, standards bodies and relevant stakeholders, including industry and civil society.

Moreover, the legal regime should govern all general purpose AI systems, which include foundation models and generative AI.

JUSTIFICATION

The GPT-3 language model exceeded expectations by not only generating text but also by learning to perform 3-digit calculations and other new tasks after the training phase had been completed. **The unpredicted gains in capability by this technology highlight the need for ongoing monitoring and potential adjustments to the legal framework governing GPAI systems**, especially as they get larger and more sophisticated, with potentially greater risks. By allowing for such oversight, the European Parliament's text provides a mechanism that ensures the AI Act remains relevant.

SUGGESTION

FLI also supports the European Parliament's proposals that require the European Commission to consult the proposed AI Office on circumstances that an AI system's output would amount to significant risk (Article 6.2); on amendments to Annex III (Article 7.2a); on guidelines for implementing the regulation, particularly when accounting for the needs of SMEs (Article 82b); on harmonised standards (Article 40.1a); common specifications (Article 41.1a and 41.2); and delegated acts amending the conformity assessments (Article. 43.5 and 48.5), including possibly subjecting other high-risk AI systems beyond biometrics to third-party assessment (Article 43.6); on the delegated acts detailing the modalities of the sandboxes (Article 53a.1); and on the annual reports, to be made public, arising from the sandboxes detailing incidents and lessons learned (Article 53.5b).

These provisions could be further strengthened by including consultation with all relevant stakeholders, including industry and civil society. More than previous technological revolutions, AI requires a uniquely close collaboration between research and regulation, leveraging synergies between technical understanding and policy expertise.