‘FUTURE-PROOF AI ACT: TRUSTWORTHY GENERAL PURPOSE AI’

Liberties’ Submission to the European Commission’s Multi-Stakeholder Consultation

September 2024

CIVIL
LIBERTIES
UNION FOR
EUROPE

Table of Contents

Executive Summary

Liberties is pleased to participate in the European Commission's public consultation on general-purpose AI in the context of the AI Act. The AI Act is a landmark piece of legislation that ushers in necessary regulatory framework and fundamental rights protections, despite several important safeguards for rights and values not making it into the adopted law. There is still important work to be done in order to ensure that those safeguards that were included are able to have the envisaged effect, and this is particularly true in the critical area of general-purpose AI (GPAI).

In this submission, we focus on certain key issues applicable to GPAI models where enforcement is critical to the protection of fundamental rights and democratic values. In particular, transparency is key not only to oversee the development and deployment of GPAI, but also to increase trust in the decision-making of these systems. The AI Act requires providers of high-risk systems to disclose information about data and data governance, technical documentation, record-keeping, transparency and provision of information to the deployer, human oversight, and accuracy, robustness and cybersecurity. Making the whole process, from development through input and outputs, this transparent for GPAI – not only high-risk systems – would make both the systems and their deployers accountable.

This level of transparency and the involvement of human oversight and supervision will raise the public's trust in technology and also ensure that government agencies, courts and authorities can rely on GPAI systems. Through reinforcement learning from human feedback, GPAI outputs are rated by humans for correctness in order to avoid false and misleading information or bias generated by GPAI. Human supervision is critically important to correct outputs that could impact fundamental rights and the rule of law. We must also have diligent oversight and transparency over GPAI systems to protect copyright and ensure that personal data are used lawfully, as they play an outsized role in training large models.

Indeed, the proper use of both personal and synthetic data is central to the protection of fundamental rights and values. It is critical to assess the nature and the amount of personal data used. The rules of the General Data Protection Regulation (GDPR) are applicable in the enforcement of the AI Act. The use of synthetic data could potentially be a substitute for personal data, providing a privacy-friendly solution; however, it must be monitored for, among other things, ensuring that it does not contribute to unfair or biased decisions.

Particular attention must also be paid to systemic risk that arises from GPAI. This is especially true because the AI Act has far too many loopholes and weak standards to fully defend fundamental rights and the rule of law. It fails to ban some uses of AI that Liberties has been advocating around, even when said uses have already been found to violate human dignity, freedom, equality, democracy, the rule of law or fundamental rights.

It is our hope that this submission draws attention to critical issues that must be addressed and helps inform future action to ensure that the AI Act and its delegated acts, including the Codes of Practice, are as strong as they can be.

Introduction

Article 3 (63) of the AI Act defines a general-purpose AI model as:

"[A]n AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market".

General-purpose AI (GPAI) systems have a wide range of possible uses, both intended and unintended by the developers. They can be applied to many different tasks in various fields, often without substantial modification and fine-tuning. These systems are becoming increasingly useful commercially due to the growing amounts of computational resources available to developers and innovative methods to use them. Current GPAI systems are characterized by their scale (memory, data and powerful hardware) as well as their reliance on transfer learning (applying knowledge from one task to another).[1]

Considering the scope of the above definition, and how diverse GPAI systems are depending on the manner of deployment, elaborating a Code of Practice for GPAI will require granular regulatory efforts and regular revision.

Our primary focus in this submission is on basic safeguards, applicable to all GPAIs, to ensure that the fundamental rights guaranteed under the AI Act, the General Data Protection Regulation, the Digital Services Act and the Directive on Copyright in the Digital Single Market are fully protected, and all related rights are able to thrive.

General-purpose AI Code of Practice

The Code of Practice on GPAI, as outlined in Article 56 of the AI Act, will establish the foundation for compliance with the rest of the AI Act, helping providers and others ensure that AI systems are designed, developed and used in uniform, conforming ways. Furthermore, the Code should be an essential component for ensuring transparency, personal data protection, copyright rules, and risk identification, assessment, management and mitigation, to ensure the protection of fundamental rights. Beyond this, the Code of Practice should also target bias and ethical considerations, like issues of fairness stemming from AI's use. Therefore, the AI Office should invite civil society organizations, academia and independent experts to

[1] Bommosani et al, On the Opportunities and Risks of Foundation Models, Stanford, 2022. https://arxiv.org/pdf/2108.07258

include their knowledge in the field and participate in the process.

The need for transparency

An underlying problem with AI systems, including the GPAIs, is the very often unknown existence of biases and discrimination, due to the opacity of the systems. Furthermore, AI developers often argue that it is impossible to put transparency into practice and that it would be difficult to do — for example, what information about which aspects of an AI system should be disclosed in order for it to be considered transparent. There are also arguments to limit transparency to protect privacy, national security, or business and other interests.[2] Liberties strongly disagrees with these arguments and believes that the carve-outs from transparency rules must be kept to a minimum, be proportionate, and any such decision should be overseen by authorities and subject to court review.

Liberties is of the opinion that transparency is key not only to oversee the development and deployment of GPAI but also to increase trust in the decision-making of these systems. The AI Act requires providers of high-risk systems to disclose information about data and data governance (Article 10), technical documentation (Article 11), record-keeping (Article 12), transparency and provision of information to deployer (Article 13), human oversight (Article 14), and accuracy, robustness and cybersecurity (Article 15). Even though the AI Act sets these requirements for high-risk AI systems, we advise requiring the same for GPAI systems above a certain size or use, including developers, deployers, providers, and operators.

Making the whole process transparent, from development through input and outputs, would create accountability for deploying these systems. The AI Act sets an obligation to disclose information about data used to train and validate GPAI. To facilitate transparency, providers are expected to produce and publish summaries of content/data used for training. Furthermore, they must provide narrative explanations.

The reason for having this transparency is to facilitate enforcement and establish accountability. Transparency over AI training data and data sources is essential for accountability in AI development and deployment. The disclosure of data sets must be a balancing exercise. We must emphasize that national security implications of trade secrets should never serve as a "blanket justification for not disclosing information about the content used to train GPAI in a situation where there are legitimate

[2] Woudstra, Fenna. What does Transparent AI mean? AI Policy Exchange, 2020. https://aipolicyexchange.org/2020/05/09/what-does-transparent-ai-mean/

reasons to make information about the training data public."[3]

The guiding principle must be transparency, while certain datasets could be summarized or some information removed to protect other interests.

The need for human supervision

Trust on behalf of operators, deployers, and impacted people depends on transparency, as discussed above, but also on the involvement of humans in oversight and supervision. Although unsupervised learning is a method that has been in use for some time, "human supervision has recently made a comeback and is now helping to drive large language models forward. AI developers are increasingly using supervised learning to shape our interactions with generative models and their powerful embedded representations."[4]

Reinforcement learning from human feedback (RLHF) has been used by many developers, among others OpenAI for ChatGPT. GPAI outputs are rated by humans for correctness in order to avoid false and misleading information or bias generated by GPAI. Human supervision is critically important to correct outputs, especially when fundamental rights and the rule of law are in question.

Training models with RLHF[5] is a potential means to ensure that GPAI aligns with human values, and fixing biased datasets and classifiers is another great advantage of RLHF. It is of the utmost importance for the process to be transparent. Accurate and reliable AI models should be able to explain the decision-making processes and the specific roles humans played in RLHF.

Copyright

GPAI models are trained on vast amounts of data, including copyrighted work. Articles 3 and 4 of the Text and Data Mining Directive (2019/790) (DTMD)[6] on copyright and related rights in the Digital Single Market created the legal basis to use copyrighted content for training. We believe that the TDMD creates the legal basis for researchers at academic research institutions and cultural heritage institutions to "use all lawfully accessible works (e.g., the

[3] Warso, Zuzana et al. Sufficiently Detailed? A proposal for implementing the AI Act's training data transparency requirement for GPAI 2024. https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf

[4] Martineau, Kim. What is generative AI? IBM, 2023. https://research.ibm.com/blog/what-is-generative-AI

[5] Lambert, Nathan et. al. Illustrating Reinforcement Learning From Human Feedback. Hugging Face, 2022. https://huggingface.co/blog/rlhf

[6] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. https://eur-lex.europa.eu/eli/dir/2019/790/oj

entire public Internet) to train ML (Machine Learning) applications. Everyone else – including commercial ML developers – can only use works that are lawfully accessible and for which their rightholders have not explicitly reserved use for TDM purposes."[7]

This means that the TDMD, along with the copyright regulatory framework, creates the legal basis for rightholders to opt-out in order to prevent their works from being used for training and establish negotiation for licensing. The opt-out regime also provides certainty for all participants in the value chain.

In order to establish proper accountability for training and for the output of GPAI models, developers must be transparent by disclosing the sources of content (works and data) used for training, including copyright-protected works. This means disclosing not only the fact that the material was used, but also how it was used.

Disclosing those data sets would improve the opportunities for rightholders to exercise their rights, supporting the informed choice of rightholders, including opting-out from the system.[8] The Code of Practice could serve as a legal basis to implement machine-readable standards to easily detect opted-out content.

Personal data

Transparency around the use of personal data sets is essential for data subjects to exercise their rights effectively. Accurate information about personal data use for GPAI systems is required by the GDPR, further supporting the core requirements of personal data protection principles, such as purpose limitation and data minimization. In order to achieve proper enforcement in relation to personal data usage to train, test or validate GPAI, we need accurate information, therefore, summaries of data sources are only a first layer that must be clearly defined and further detailed. A sufficient level of summary should provide meaningful transparency, which should include informing the data subject in case their personal data was used for training, testing or validating the AI system.

It has been a general argument on behalf of GPAI providers that they work with a 'black box' and are thus unable to identify personal data used for training. The European Data Protection Board,[9] however, clearly stated that

[7] Communia, Policy paper #15 on using copyright works for teaching the machine. 2023. https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/

[8] Sufficiently Detailed proposal for implementing the AI Act's training data transparency requirement for GPAI, Open Future, 2024. https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf

[9] Report of the work undertaken by the ChatGPT Taskforce, European Data Protection Board, 2024. https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf

this fact cannot be invoked as a reason for non-compliance.

With regard to GPAI, there are two territories of personal data use that are worth further investigating, because they fundamentally differ from other processing mechanisms by software. One is the input data that could be based on web scraping personal data from publicly accessible sources such as websites. In these cases, Article 14 of the GDPR applies. The other territory of personal data use is when personal data is part of the output data. Even though it could potentially contain significant valid or invalid personal data depending on the output of the GPAI system, we must mention that even invalid personal data is personal data, and the data subject has the right to rectify. In this case the principle of data accuracy also plays a role.

It is critical to design a system with thorough data-protection safeguards. Data minimization is one of the core principles of the GDPR.

Synthetic data

There are certain situations when the lack of real data, or the opportunity to use those data, could be substituted by synthetic data, which is generated to augment or replace real data to improve AI models, protect personal data, and mitigate bias. The positive aspect is that synthetic data is labeled and could potentially improve accuracy and reliability without infringing privacy, data protection or copyrights. But we must warn that synthetic data is not a silver bullet. Depending on the generation techniques and data sources, it can create mistakes and unfair decisions, and therefore scrutiny is still needed in relation to the use of synthetic data. The ultimate aim is to generate data that can't be traced to individuals but yet preserve the statistical properties of the original data.10

Synthetic data could also be used for testing the model for flaws and biases, to show where the AI model is likely to make mistakes or biased decisions – "they can help to make AI models more fair, accurate, and trustworthy."11

However, the application of synthetic data comes with challenges. It can easily amplify biases in the synthetic data set. The reliability of synthetic data is also questionable, because it misses the real-world nuances of personal data. Model collapse is also a possible scenario when a system relies too heavily on synthetic data. Safeguarding rigorous testing, monitoring and refinement are essential to substitute real personal data with synthetic data.

Enhancing trust in AI is interconnected with transparency. There is a broad agreement that

[10] Martineau, Kim and Feris, Rogerio. What is synthetic data? IBM, 2023. https://research.ibm.com/blog/what-is-synthetic-data

[11] IBM Research, Five ways IBM is using synthetic data to improve AI models https://research.ibm.com/blog/synthetic-data-explained

feeding AI models with high-quality data is a basic requirement to prevent problems later. However, data-cleansing techniques that filter data for bias, hate, or private information before training must be used by companies. Information about data-cleansing standards must also be revealed for accountability reasons.

Systemic risk

GPAI models have become extensively integrated into both public and private areas of life. High-impact GPAI models are almost exclusively developed by a few companies, dominated by Big Tech. This presents a possible systemic risk of economic power increasingly being centralized in the hands of a few actors with an outsized degree of control over access to this technology and its economic benefits, perhaps exacerbating inequalities between countries. Furthermore, as developers input certain values and principles into GPAI models, this risks centralization of ideological power, producing models that are not fit to adapt to evolving and diverse social views or that create echo chambers. Finally, too-rapid adoption of this technology could outpace the ability of society to adapt, straining the labor market, education system and public discourse, among other things.[12]

Expanding on the inclusion of systemic risk in the AI Act, it will be necessary to develop a taxonomy of the systemic risks of GPAI and map existing GPAI models that present these characteristics.[13] The AI Office is responsible for overseeing GPAI and will determine when a system has a "significant impact" on the market, but this consideration should reflect national-level market parameters and impact as well. If, for example, a system has a significant impact on the Estonian market but not elsewhere, the Office should consider that this system indeed has a significant impact regardless of its confinement to one domestic market. To aid work around systemic risk, the EU should adopt a public-interest-focused Code of Practice on GPAI and push for the designation of systems that pose systemic risk.

The outputs of generative AI systems must be marked and detectable as artificially generated. Given the high public profile of GPAI, the ability of the AI Act to have an impact on these systems will be a key test of its effectiveness. Civil society should be part of the Code of Practice working group organized by the AI Office, to help define these obligations, identifying and adjusting requirements to create the possibility to pose systemic risk.

[12] Maham, Pegah and Kuspert, Sabrina. Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks. Interface, 2023. https://www.interface-eu.org/storage/archive/files/snv_governing_general_purpose_ai_pdf.pdf

[13] Iwanska, Karolina et. al. Towards an AI Act that serves people and society. European Center for Not-for-Profit Law (ECNL), 2024. https://ecnl.org/sites/default/files/2024-08/AIAct%20implementation_ECNL%20report.pdf

As already flagged by civil society in advocacy around the Digital Services Act, identifying "systemic risk" is not merely conceptual but practical — practical problems could include a lack of frequent and transparent information about the risk assessment work already performed by platforms, and uncertainty about how the necessary "learning by doing" will be conducted in an effective and collaborative fashion. These outstanding concerns around the DSA must be similarly considered under the AI Act and by the Office.[14]

Fundamental rights and rule of law

Liberties is of the opinion that the AI Act fails to effectively protect the rule of law and civic space, instead prioritizing industry interests and those of security services and law enforcement bodies. While the AI Act requires developers to maintain high standards for the technical development of AI systems (e.g., in terms of documentation or data quality), measures intended to protect fundamental rights, including key civic rights and freedoms, are insufficient to prevent abuses. They are riddled with far-reaching exceptions, lowering protection standards, especially in the area of law enforcement and migration.

The AI Act introduces prohibitions, but they are rife with loopholes, which calls into question how effective they will be in protecting civic space and fundamental rights. The AI Act also fails to ban some uses of AI, and creates carve-outs for national security purposes, even when they have already been found to violate human dignity, freedom, equality, democracy, the rule of law or fundamental rights. These include: real-time remote biometric identification in public spaces (e.g. face recognition) in the area of law enforcement (with vast exceptions); biometric categorisation to infer sensitive information about people (e.g. their race or sexuality), with a blanket exception for law enforcement; creating or expanding facial recognition databases through scraping of facial images from the internet or video surveillance footage; emotion recognition in education or employment; predictive policing when it is based on profiling individuals (as opposed to predicting crime based on criminal statistics from a certain neighborhood) and only when it is not supporting an assessment by a police officer.

It has been documented how AI-driven technologies are used to surveil human rights activists, journalists, assess whether air passengers pose a terrorism risk, and appoint judges to court cases. Rule of law and fundamental rights standards require robust safeguards to protect people and societies from the negative impacts of AI — to insulate from its abuses our judiciary and legal system, our elections and democratic processes, and our fundamental rights. The AI Act has far too many loopholes and weak standards; none of the

[14] Marsch, Oliver. Researching Systemic Risks under the Digital Services Act. AlgorithmWatch, 2024. https://algorithmwatch.org/en/researching-systemic-risks-under-the-digital-services-act/

aforementioned areas is properly protected. It is therefore imperative that the European Commission and other bodies responsible for the delegated acts, the implementation and enforcement of the AI Act proactively facilitate civil society participation and prioritize diverse voices, including those of people affected by various AI systems.[15]

Conclusion

The AI Act and the upcoming delegated acts, along with standards for AI models, will sweep in rules, standards and Codes of Practice on artificial intelligence development, use, and governance that are sorely needed. The AI Office and the Board must take crucial steps to ensure that the AI Act lives up to its promises on fundamental rights and the rule of law. The Code of Practice, as outlined in Article 56 of the AI Act, will be an essential component for ensuring transparency, personal data protection, copyright rules, and risk identification, assessment, management and mitigation. The Code of Practice will establish the foundation for compliance with the rest of the AI Act, helping GPAI providers and others ensure that AI systems are designed, developed and used in uniform, conforming ways, respecting fundamental rights.

[15] Day, Jonathan et. al. Packed with loopholes: Why the AI Act fails to protect civic space and the rule of law. Civil Liberties Union for European, European Center for Not-for-Profit Law, European Civic Forum, 2024. https://dq4n3btxmr8c9.cloudfront.net/files/hjoz6a/AI_Act_RoL_Analysis.pdf

The Civil Liberties Union for Europe (Liberties) is a non-governmental organisation promoting the civil liberties of everyone in the European Union. We are headquartered in Berlin and have a presence in Brussels. Liberties is built on a network of 19 national civil liberties NGOs from across the EU.

Publisher
Civil Liberties Union for Europe e.V
Ebertstraße 2. 4th floor
10117 Berlin , Germany
www.liberties.eu

Authors
Jonathan Day jday@liberties.eu
Eva Simon eva.simon@liberties.eu