ALLAI.

Stellungnahme für den Bundestagsausschuss für Digitales

Öffentliche Anhörung „Generative Künstliche Intelligenz“
am Mittwoch, 24. Mai 2023, 14:30 – 16:30 Uhr

Catelijne Muller, ALLAI

Deutscher Bundestag
Ausschuss für Digitales

Ausschussdrucksache
20(23)158

24.05.2023

1. General Remarks

So-called general purpose AI-systems (GPAI) have become a contentious issue in the ongoing legislative process of the European Artificial Intelligence Act (AIA). Different proposals have been floated, ranging from completely excluding GPAI from the scope of the AIA, to establishing a separate status for them.

Also, the recent developments around Large Language Models such as GPT- 3.5 and GPT-4 underpinning ChatGPT, AutoGPT and BabyAGI, including several open letters and an investigation into ChatGPT by the Italian data protection supervisor, have put the topic of GPAI and also Generative AI (GenAI) and Foundation Models (FM) in an even more critical light.

1.1 What are General Purpose AI, Generative AI and Foundation Models?

First and foremost, it remains under discussion what exactly is meant by 'general purpose AI' and where the differences lie between GPAI, Generative AI and Foundation Models. The European Parliament recently proposed two definitions, one for General Purpose AI and one for Foundation Models. There is also the element of generative AI (GenAI), that further confuses things. The Member States have adopted their position on the AIA in December, defining GPAI as an AI system that *"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."* As far as we can see, the European Parliament considers Foundation Models a subset of GPAI, and GenAI a functionality. In this reaction we will reply to these three terminologies jointly as GPAI/GenAI/FM.

1.2 Singular points of failure with broad impact

There is an obvious trend towards ever fewer, more general and often very large models. While these models have demonstrated impressive behavior, they can also fail unexpectedly (hallucinate), harbor biases, and are poorly understood. As these systems are

deployed at scale, they can become singular points of failure that radiate harms (e.g., security risks, discrimination, inequities) to countless downstream AI applications.

That is not to mention the multiple legal and ethical issues these models present, such as around data protection, IP rights, automation bias, manipulative power, the scaling of misinformation, skills erosion, potential job displacement, the risk of uncontrollable autonomy, and so on.

1.3 Benchmark datasets

Apart from 'general' AI-models, there is a wide practice of using so-called 'benchmark' datasets that form the backbone of machine learning research and development. Recent critical inquiry into these datasets have however revealed biases, poor categorization and offensive labeling in these datasets. Even many of the fairness in ML researchers use datasets 'as is' without checking them for completeness, representativeness and overall fairness.

1.4 Homogeneity

The issues described above around GPAI/GenAI/FM (consisting of ever fewer and more general models and benchmark datasets) can be referred to as the 'homogeneity problem'. Machine learning by its nature results in more homogeneous decision making compared to human decisions. If ever fewer machines inform ever more decisions, biases and errors could become amplified and embedded and generalized throughout society.

1.5 GPAI/GenAI/FM & the AIA - intended purpose *and* reasonably foreseeable use

One of the arguments for creating a separate status for GPAI/GenAI/FM in the AIA is that GPAI/GenAI/FM providers do not know for which purpose their system will be used, so the risk category of their system cannot be determined up front.

In our paper AIA in-depth #1 | Objective, Scope, Definition we propose an approach to tackle this, which is common in Union legislation regarding product safety. This approach is to add the notion of '*reasonably foreseeable use*'. Given the potential impact of these GPAI/GenAI/FM systems, it is not unreasonable to ask from their providers to try to foresee the potential uses of their system and categorize their system accordingly. In other words, if it is reasonably foreseeable that a GPAI/GenAI/FM system will be used as (part of) a high risk AI system as listed in ANNEX II or III of the AIA, then the GPAI/GenAI/FM system itself classifies as high risk. Appreciating that not all uses can be foreseen, the notion would cover only those uses that are *reasonably* foreseeable.

The Member States' General Approach incorporates a new chapter on General Purpose AI, including this notion of 'foreseeable use' albeit in a slightly different manner in two parts of art. 4b:

1. *General purpose AI systems which may be used as high risk AI systems or as components of high risk AI systems (...)*

6. In complying with the requirements and obligation referred to in (...):
- any reference to the intended purpose shall be understood as referring possible use of the general purpose AI systems as high risk AI systems or as components of AI high risk systems in the meaning of Article 6;

It also proposed that specific requirements for GPAI/GenAI/FM should be set by the European Commission at a later stage. For more on the Council position on GPAI/GenAI/FM, we refer to our AIA Policy Analysis | Council General Approach.

1.6 GPAI/GenAI/FM should be held to a higher standard

Another argument for creating a separate status for GPAI/GenAI/FM is that these systems always need to be 're- or uptrained' before they can be used for a certain purpose in a new domain (think of tumor detection in healthcare). Hence, the GPAI/GenAI/FM provider, in its compliance process, could never 'anticipate' the multitude of downstream applications that would go through such re-training process.

First, the data requirements of the AIA already deal with this issue in a clever way. Paragraph 2(g) allows for "*the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed*" (see ANNEX I to this paper). That leaves the responsibility of delivering a high quality, robust and trustworthy core functionality with the GPAI/GenAI/FM provider, including the obligation to properly inform any downstream user of possible data gaps or shortcomings in high risk use cases.

For GPAI/GenAI/FM one could even argue that because of their potential use in a wide variety of high-risk domains (healthcare, critical infrastructure, law enforcement), they should be held to a *higher* standard instead of a lower one. In fact, the overall Union objective of safety and liability legal frameworks, is to ensure that all products and services, including those integrating emerging digital technologies, operate safely, reliably and consistently and that damage is remedied efficiently. The EU follows a different approach than other parts of the world, where responsibility is determined afterwards, often leading to large liability claims. It would also break with the overall objective of the AIA which is to protect health, safety and fundamental rights from adverse effects of AI.

1.7 Shifting responsibility downstream will stifle innovation

Limiting the scope of the AIA for GPAI/GenAI/FM, also runs the risk of in fact stifling innovation rather than supporting it. Setting less requirements or lower standards at GPAI/GenAI/FM provider level, would shift the responsibility of bringing these systems in compliance with the AIA to 'downstream' users. They would be the ones having to comply with the requirements for high risk AI, which might be too much of a burden, especially for SME's and micro-enterprises, or perhaps even technically impossible.

Even if the GPAI/GenAI/FM developer would help 'downstream users' with the technicalities of complying with the AIA, it places the latter in a fully dependent position. As a result, this could lead to a limited uptake of GPAI/GenAI/FM systems on the one hand, and a (further) concentration of AI innovation power with GPAI/GenAI/FM developers on the other.

To avoid GPAI/GenAI/FM developers limiting or even excluding also their *liability* vis-a-vis downstream users in contracts or terms and conditions, the EP already proposed language to avoid these types of contractual provisions.

1.8 Requirements for high risk AI and GPAI/GenAI/FM

We have not yet seen any overview that indicates which requirements are in need of adaptation for GPAI/GenAI/FM systems or cannot be fulfilled by GPAI/GenAI/FM providers. For that reason, we preliminarily assessed the requirements in light of GPAI/GenAI/FM in ANNEX I to this paper. We added the earlier mentioned notion of 'reasonably foreseeable use' to the requirements, meaning that each requirement is seen in light of the reasonably foreseeable use of the GPAI/GenAI/FM system.

This preliminary assessment indicates that there are only a few elements of the requirements for high risk AI (Chapter 2 of Title III AIA) that would be theoretically difficult for GPAI/GenAI/FM providers to meet due to the fact that they do not know how their system will be used. In fact, a number of requirements can only be met (i.e. built into the system) by the GPAI/GenAI/FM provider, and not by the downstream user. We emphasize that we did not consider whether it is generally possible to meet the requirements. In fact, if not, the system will not comply with the AIA no matter who is responsible for it.

This, compared to the full responsibility for downstream providers of having to meet all the requirements, provides a strong argument for having the current requirements apply to GPAI/GenAI/FM providers as well. Given the recent developments around generative AI, additional requirements might be necessary, especially around IP rights, manipulation, machine autonomy and potential emergent behavior.

1.9 Liability

The recent proposal for an AI Liability Directive (in combination with the proposal for a revision of the Product Liability Directive) makes it even more pertinent to include GPAI/GenAI/FM in the AIA. These proposals consider non-compliance with the AIA cause for the presumption of causality between the provider and the AI system. Excluding GPAI/GenAI/FM providers from the scope of the AIA would thus also bring them beyond the reach of the AI Liability Directive as well.

1.10 ChatGPT, AutoGPT, BabyAGI

Large Language Models have taken the world by storm in the past couple of months. Much has already been said about them and their risks have been listed extensively. OpenAI itself

has described (and tested) potential risks in its GPT-4 system card. The model exhibits the tendency to 'hallucinate' (i.e. provide wrong information, including non-existing scientific papers, false accusations, incorrect calculations and so on). An Australian Mayor has sued OpenAI for ChatGPT wrongfully accusing him of bribery and having spent time in prison. Experts warn that the internet could be flooded with fake news and polarizing content. Europol has warned of an increase in criminal activities such as hacking, cyberattacks and phishing, that can become far easier with the help of ChatGPT. Teachers are struggling with students having their homework done by ChatGPT. Companies are prohibiting the use of ChatGPT by their workforce as it jeopardizes their business model. And so on.

1.11 Separate requirements for GPAI/GenAI/FM

OpenAI also describes the potential risk of 'autonomous replication'. While their tests found that GPT-4 (the Large Language Model underpinning ChatGPT) was ineffective at the autonomous replication task based on preliminary experiments, they note that additional tests are necessary to come to a reliable judgment of risky emergent capabilities. of GPT-4.

In the last couple of weeks, we have however seen experiments showing some form of autonomous replication. Computer scientists built several applications on top of GPT-4, the most notable being AutoGPT and BabyAGI, that are able to generate and execute consequent tasks themselves, based on only one human defined 'goal'. These systems are capable of searching the internet, opening a google account, setting up a google drive folder, opening a file and adding text to that file, without the need for additional human intervention. In particular BabyAGI has shown a form of autonomous replication, where it split a human given goal up into several subtasks, that were then executed simultaneously by different GPT-4 language models it initiated itself. It should be noted that the computer scientists themselves acknowledge that safeguards need to be put in place for these systems.

1.12 AI-driven manipulation

Recently, a Belgian man committed suicide after a lengthy conversation with a chatbot running on a Large Language Model. According to his wife, the conversation with the chatbot took a disturbing turn and led to the man's suicide. Another company, exploiting a chatbot-app establishing intimate relations with users, found their users becoming mentally distressed after it had toned down the level of intimacy of the conversations. In a reaction it added the number of the suicide hotline to the app.

The powerful effects of AI-manipulation, including those embedded in chatbots, are currently not sufficiently understood or addressed and cannot be curbed by merely imposing transparency measures. In our paper AIA in-depth #2 | Prohibited AI Practices, we argued that the AIA provides a grand opportunity to address the legal gaps and the wider societal harms that AI-driven manipulation can bring. A prohibition of AI-practices aimed at, or resulting in, deception, material distortion of behavior or exploitation of a person's vulnerabilities would fit well within the larger objective of the AIA. We proposed amending the prohibition of art. 5 (a) and (b), which has already been partially taken up by the Council.

We acknowledge that enforcing this prohibition will be a challenge, but legislation holds many enforceability challenges. That has not stopped us from regulating before. A clear prohibition like this will on the other hand have a great preventive effect, that should not be underestimated.

We realize that this could mean that GPAI/GenAI/FM systems will always have to comply with the requirements for high-risk AI, even if they are used in low risk domains or applications. We do argue however that compliance with them will lift the quality, reliability and trustworthiness of GPAI/GenAI/FM systems in general, setting a positive trend overall.

Given the recent developments around generative AI and in particular Large Language Models, additional safeguards/requirements for GPAI/GenAI/FM might even be necessary.

2. Answers to questions of the Bundestagsausschuss für Digitales

1) The regulation of generative AI is currently the subject of negotiations in the context of the European AI Act (AIA). In your view, how can generative AI be effectively incorporated into and regulated by the AIA, and what is your opinion of the proposed distinctions, within generative AI, between "general purpose AI" and "foundation models"?

See general remarks.

2) Generative AI offers many potential uses in a wide range of different occupations, and can alleviate pressure on the labour market. What is your view of the potential and the risks of generative AI for the world of work, and in what areas do you believe regulation is needed?

Rapid technological changes have historically often resulted in loss of jobs but also in the emergence of new jobs. Yet, the question is with what time lag the loss and creation of jobs happens and how the benefits of the new created value is distributed. With the current trajectory of generative AI technology, job losses could happen very quickly. For instance, many copywriters have recently been asked just to review AI generated texts or have been laid off entirely.

3) To what extent can applications from government or economic systems which do not always share democratic and liberal values have an impact on European society, and how should the EU and Germany deal with this?

The AI Act will have a global reach as it will cover all systems that affect EU citizens, no matter where they are being developed. Similarly as with the GDPR, it will have extraterritorial impact.

5) Many proposals are currently circulating on how to appropriately address the regulatory challenges of generative AI applications in the EU proposals for an AI Regulation and an AI Liability Directive. Is a risk-based approach even suitable for the regulation of generative AI, or is a systematic risk analysis needed, for example, similar to the risk analysis and minimisation mechanism contained in the DSA?

Both the risk based approach (see answer to question 1) and a systemic risk analysis could be applied simultaneously. Moreover, the requirements regarding a risk management system and post market monitoring of the AI Act are also meant to analyze systemic risks. The European Parliament proposes to add the protection of democracy, rule of law and the environment as the objective of the AIA (which includes systemic risks).

6) Should new phenomena and issues be expected in terms of generative AI applications having a negative influence on the democratic opinion-forming process? How can media freedom and diversity of opinion be strengthened in legal and political terms in the age of generative AI, including – but not only – with regard to appropriate remuneration for journalists, artists and creative professionals? Where do you believe adjustments may be necessary, for example in copyright law?

One of the main risks of GenAI is the proliferation of misinformation and deep fakes at scale which can give rise to more filter bubbles and proliferation of fake news, disinformation and propaganda, and affects the capacity of individuals to form and develop opinions, receive and impart information and ideas and thus impact our freedom of expression and the functioning of democracy.

14) In your view, what rules on generative AI are needed in the AI Act, specifically with regard to obligations for developers of foundation models to pass on information within the supply chain? What are the advantages and disadvantages of such obligations? Above what threshold should the high-risk rules contained in the AI Act apply to applications based on generative AI?

I welcome the obligations proposed by the IMCO/LIBE Committees of the European Parliament in the new art. 28b of the AIA, stating that providers of foundation models shall:

(a) demonstrate through appropriate design, testing and analysis that the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development with appropriate methods such as with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development;
(b) process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation;
c) design and develop the foundation model in order to achieve throughout its lifecycle appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity assessed through appropriate methods such as model evaluation with the involvement of independent experts, documented analysis, and extensive testing during

conceptualisation, design, and development;
(d) design and develop the foundation model, making use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency, and the overall efficiency of the system. This shall be without prejudice to relevant existing Union and national law and this obligation shall not apply before the standards referred to in Article 40 are published. They shall be designed with capabilities enabling the measurement and logging of the consumption of energy and resources, and, where technically feasible, other environmental impact the deployment and use of the systems may have over their entire lifecycle;
(e) draw up extensive technical documentation and intelligible instructions for use in order to enable the downstream providers to comply with their obligations pursuant to Articles 16 and 28.1.;
(f) establish a quality management system to ensure and document compliance with this Article, with the possibility to experiment in fulfilling this requirement;
(g) register that foundation model in the EU database referred to in Article 60, in accordance with the instructions outlined in Annex VIII paragraph C;
(h) for a period ending 10 years after their foundation models have been placed on the market or put into service, keep the technical documentation referred to in paragraph 1(c) at the disposal of the national competent authorities;

I also welcome the obligation for providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI") and providers who specialize a foundation model into a generative AI system to:
a) comply with the transparency obligations outlined in Article 52 (1);
b) 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 in line with the generally- acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression;

c) without prejudice to national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

17) To what extent does the distribution of the advantages and disadvantages of general purpose AI vary for different population groups (both within national societies and from a global perspective, in terms of the Global South/North) as a result of the factors listed below:
    - Differences in access to technology (e.g. because of different technical, material, educational or other conditions)
    - Differences in representation in training data (e.g. health data of women versus men, and of white people versus people of color, African languages versus English, etc.)
    - Differences in the degree to which people are affected by stereotypical associations and discrimination (e.g. on the basis of gender or ethnicity)

- Differences in the burden imposed by the resource consumption caused by AI systems and how could a fairer distribution of the advantages and disadvantages be achieved?

As with all other technologies, general purpose AI is shaped by the societies in which it is developed and tends to replicate structures of discrimination and oppression of these societies. General purpose AI is mostly developed by white, educated, male programmers in the capitalist interest of shareholders. The interests of ecosystems or systematically disadvantaged groups are not taken into account equally in this process. This often results in AI discriminating against disadvantaged groups and many clickworkers in the Global South are exploited in the process of labeling data for and training AI models. Technologies such as general purpose AI do have an emancipatory potential to free people from unnecessary and unrewarding toil, yet in the current trajectory it is not realizing this potential. Currently, the benefits of AI currently mostly accrue to those with economic and political power.

18) Should generative AI, as a multi-purpose AI, generally be classed as a high-risk AI within the meaning of the European AI Regulation, resulting in it having to meet higher standards? How sensible/feasible do you believe the various regulatory options for generative AI are, such as transparency obligations regarding training data and training processes, an obligation for providers of a general purpose AI to perform and publish a risk assessment, visible or imperceptible marking of all or certain AI-generated content, the right to verifiable freedom from discrimination and to access for researchers, and other options under discussion?

See general remarks.

ANNEX I: Preliminary assessment of the requirements for High Risk AI in light of GPAI/GenAI/FM-systems*

Article 9 (Risk management system)
The risk management system as described in art. 9, being a continuous iterative process of detecting risks to health, safety and fundamental rights, seems to be a fairly reasonable system also for GPAI/GenAI/FM providers to be set up. Such a system would describe how the risks of the GPAI/GenAI/FM system in question are managed, in particular where these risks can affect (via API) downstream applications.

Article 10 (Data and data governance)
Many if not all current 'GPAI/GenAI/FM-systems' are data-driven, so the requirement for proper data governance seems to be crucial here. Some notable elements: Paragraph 2(g) allows for "*the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed*", which could solve the 're- and uptraining' issue (where a system needs re- or uptraining for a particular purpose), as mentioned above).

Paragraph 3, setting requirements for training, validation and testing data, has two parts. A general part, which requires that "*training, validation and testing data sets shall be relevant, representative and [to the best extent possible,] free of errors and complete* [and] *They shall have the appropriate statistical properties.*" And a specific part, where, if applicable, specific use cases or domains trigger a set of additional data requirements *"as regards the persons or groups of persons on which the high-risk AI system is intended to be used.*" GPAI/GenAI/FM providers could easily comply with the first part. If applicable, i.e. for the reasonably known use cases or domains of their system, they could even comply with the second part.

Paragraph 4 could be easily amended to reflect the above: *Training, validation and testing data sets shall take into account, to the extent required by the* ***reasonably known or foreseen*** *purpose.*

Article 11 (Technical documentation)
This requirement seems reasonable and even desirable given GPAI/GenAI/FM providers' responsibility vis-à-vis downstream users.

Article 12 (Record-keeping)
This requirement is aimed at designing and developing AI systems in such a way that their workings are traceable. It explicitly is not aimed at performing actual tracing activities. In other words, the system needs to technically allow for recording and logging. This is in fact one of those requirements that would be impossible to meet by downstream providers if the GPAI/GenAI/FM provider does not have these capabilities built into the system. This makes the requirement in fact very relevant for GPAI/GenAI/FM providers, particularly from a business point of view, as it would mean that GPAI/GenAI/FM systems without proper logging capabilities will not be used for high risk AI systems.

We do suggest a textual change for paragraph 4: *For high-risk AI systems referred to in paragraph 1, point (a) of Annex III, the logging capabilities shall* ***enable****, at a minimum (...):*

Article 13 (Transparency)
This requirement is aimed at designing and developing the AI system in such a way to ensure that its operation is sufficiently transparent. It requires instructions of use, change logs and technical measures to facilitate interpretation of the output of AI systems.

Exempting GPAI/GenAI/FM systems from this requirement would leave them the black boxes they often are, making it extremely difficult if not impossible for downstream providers that use GPAI/GenAI/FM systems as a component of a high risk AI system to comply with the requirement.

Two notable elements:
Almost all sub-requirements of Art. 13 can be met by GPAI/GenAI/FM providers, except for the 'human oversight measures' as described in art. 14.3(b) and referred to in paragraph art. 13.3(d).
Technical oversight measures (as described in art. 14.3(a)) can most likely only be implemented at the core of the AI system, which will be the GPAI/GenAI/FM system, and not be built in afterwards.

Article 14 (Human oversight)
This requirement does not prescribe any actual human oversight activity, but only requires that the design of the system ensures the possibility of human oversight. As described above under art. 13, technical oversight measures can most likely *only* be implemented at GPAI/GenAI/FM level.

Exempting GPAI/GenAI/FM providers from this requirement would put the burden of ensuring that the system can effectively be overseen by humans on downstream users, which may prove to be impossible if the GPAI/GenAI/FM system does not provide for that possibility.

The only element that could likely not be met by GPAI/GenAI/FM providers is the 4-eye requirement of paragraph 5.

Article 15 (Accuracy, robustness and cybersecurity)
We propose making this particular requirement a blanket requirement for all AI systems, irrespective of their risk level, in particular where it comes to cyber security. As regards GPAI/GenAI/FM systems, the requirements of accuracy and robustness, can be met also by GPAI/GenAI/FM providers, provided that the notion of 'reasonably foreseeable use' is incorporated in paragraph 1.

A notable element:
For AI systems that 'continue to learn', which can be read as 'are re-or uptrained' for a particular use, paragraph it says: "*High-risk AI systems that continue to learn after being*

placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as in an input for future operations ('feedback loops') are duly addressed with appropriate mitigation measures." As such the article already partially deals with the problem of not knowing for certain how and where the GPAI/GenAI/FM system will be used.

**This assessment does not determine whether any of the requirements can be met at all from a technical perspective. If a requirement cannot be met due to the particular technical incapabilities, the system will not comply with the AIA, no matter who is responsible for such compliance.*