ALLAI.

31.03.2023

#GPAI

AIA topics

General Purpose AI

Updated to include generative AI (GPT-4 e.o.)

This policy brief is part of a series of short analyses of important, decisive or divisive topics in the legislative process regarding the European Commission proposal for a Regulation for Artificial Intelligence (AIA).

Author:
Catelijne Muller, LL.M

GENERAL PURPOSE AI

What is the issue?

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 in an even more critical light.

As the legislative process of the AIA is entering a crucial phase, with the EP deciding on its position in the coming weeks, and the trilogues starting after that, we provide an analysis of the position of GPAI within the AIA.

What is General Purpose AI?

First and foremost, it remains under discussion what exactly is meant by 'general purpose AI'. It should in any case *not* be confused with Artificial General Intelligence, which is AI that, at a cognitive level, is as capable as humans. This type of AI does not exist.

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."*

It should be noted that pattern recognition is a (core) functionality of almost all current AI-systems. This, as well as the words *"(...) and others"* make the definition vague and multi-interpretable which leads to legal uncertainty and can create loopholes.

Singular points of failure with broad impact

There is an obvious trend towards ever fewer, very large models. While these models have demonstrated impressive behaviour, they can also fail unexpectedly (hallucinate), harbour 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. There even is a lack of agreement on basic questions such as when these models are even “safe” to be released.[1] [2]

[1] Bommasani et al (2021) “On the Opportunities and Risks of Foundation Models

[2] Future of Life Open Letter "Pause Giant AI Experiments"

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.

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 labelling[3] in these datasets. Koch et al. have found increasing concentration on fewer and fewer datasets in the field of AI research.[4] Despite widespread recognition that datasets are critical to the advancement of the field, careful dataset development is often undervalued and disincentivized, especially relative to algorithmic contributions.[5] Even many of the fairness in ML researchers use datasets ‘as is’ without checking them for completeness, representativeness and overall fairness (Propublica’s COMPAS dataset is widely used in this field while there is literature that suggests that a data processing error was made that resulted in a recidivism rate inflation of over 24%).[6]

[3] Koch et al. (2021)
[4] Ibid.
[5] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna (2021): Do datasets have politics? Disciplinary values in computer vision dataset development; Nithya Sambasivan et al. (2021): Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI.
[6] Mathias Barenstein (20-19) ProPublica’s COMPAS Data

Homogeneity

The issues described above around GPAI (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 to the point where they create structural societal drawbacks.[7]

[7] Creel and Hellman (2021): The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems, *Virginia Public Law and Legal Theory Research Paper*, no. 2021–13.

GPAI & the AIA - intended purpose *and* reasonably foreseeable use

One of the arguments for creating a separate status for GPAI in the AIA is that GPAI 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 systems, it is not unreasonable to ask from their providers to try to foresee the potential uses of their system and categorise their system accordingly. In other words, if it is reasonably foreseeable that a GPAI 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 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 should be set by the European Commission at a later stage. For more on the Council position on GPAI, we refer to our AIA Policy Analysis | Council General Approach.

GPAI systems should be held to a higher standard

Another argument for creating a separate status for GPAI 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 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 the GPAI provider, including the obligation to properly inform any downstream user of possible data gaps or shortcomings in high risk use cases.

For GPAI 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 in stead 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.

Shifting responsibility downstream will stifle innovation

Limiting the scope of the AIA for GPAI, also runs the risk of in fact stifling innovation rather than supporting it. Setting less requirements or lower standards at GPAI 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 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 systems on the one hand, and a (further) concentration of AI innovation power with GPAI developers on the other.

To avoid GPAI developers limiting or even excluding also their *liability* vis-a-vis downstream users in contracts or terms and conditions, the EP will likely propose a ban on these types of contractual provisions.

Separate requirements for GPAI

As said, the Council General Approach proposes that the European Commission sets a separate set of requirements for GPAI at a later stage.

We have not yet seen any overview that indicates which requirements are in need of adaptation for GPAI systems or cannot be fulfilled by GPAI providers. For that reason, we preliminarily assessed the requirements in light of GPAI 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 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 difficult for GPAI 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 provider, and not by the downstream user. We emphasise 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 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 behaviour.

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 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 providers from the scope of the AIA would thus also bring them beyond the reach of the AI Liability Directive as well.

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.[8] 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 polarising content. Europol has warned for and 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 jeopardises their business model. And so on.

[8] OpenAI "GPT-4 System Card (2023)

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 judgement of risky emergent capabilities. of GPT-4.

In the last couple of days, 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 show autonomous behaviour, including 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.[9] It should be noted that the computer scientists themselves acknowledge that safeguards need to be put in place for these systems.

[9] LangChain Agents Webinar

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.

Conclusion

Our overall conclusion is that GPAI systems can and should be held to at least the same standards as high-risk AI systems, if their use as (safety components of) harmonized products (ANNEX II) or in high risk domains (ANNEX III) is reasonably foreseeable, for several reasons:

- It is virtually impossible to define GPAI systems in a legally sound and definitive way, without creating legal uncertainty and loopholes.
- GPAI systems are poorly understood and can fail in unexpected ways. As such they can become singular points of failure with broad impact through numerous downstream AI systems.

- The fact that GPAI systems are likely to be used in critical (harmonized and high risk) domains makes the need for compliance with the requirements all the more pertinent.
- The trend towards ever fewer and more general AI systems leads to homogeneity of outcomes, also making the need for compliance with the requirements more pertinent.
- Less requirements or lower standards for GPAI systems places the burden of compliance with the AIA entirely on downstream users which could lower the uptake of GPAI systems (especially by start-ups and SME's) and stifle innovation.
- Less requirements or lower standards places downstream providers in a completely dependent position vis-à-vis the GPAI providers, giving the latter a competitive advantage.
- Virtually all requirements for high risk AI can be met by GPAI providers, and some of them can only be met through the GPAI design.

We realise that this could mean that GPAI 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 systems in general, setting a positive trend overall.

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

As regards the increasing ability of AI to manipulate people, these could be categorised as a prohibited AI practice under article 5. We call on the lawmakers to strengthen articles 5.1 (a) and (b) to this effect.

ANNEX I: Preliminary assessment of the requirements for High Risk AI in light of GPAI-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 providers to be set up. Such a system would describe how the risks of the GPAI 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-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 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.*
- In our paper AIA in-depth #3b | High Risk AI Requirements we argue for deletion of paragraph 5.

Article 11 (Technical documentation)
This requirement seems reasonable and even desirable given GPAI 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 the GPAI provider does not have these capabilities built into the system. This makes the requirement in fact very relevant for GPAI providers, particularly from a business point of view, as it would mean that GPAI 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 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 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 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 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 level. Exempting GPAI 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 system does not provide for that possibility.

- The only element that could likely not be met by GPAI 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 systems, the requirements of accuracy and robustness, can be met also by GPAI 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 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.*

ALLAI.

ABOUT ALLAI

ALLAI is an independent organisation that aims to foster, promote and achieve the responsible development, deployment and use of AI.

ALLAI's mission is to take a holistic approach to AI, taking into account all impact domains such as economics, ethics, privacy, laws, safety, labour, education, etc. ALLAI aims to involve all stakeholders in its mission: policy-makers, industry, social partners, consumers, NGOs, educational and care instructions, academics from various disciplines.

ALLAI was founded by the three Dutch members of the High Level Expert Group on AI, Catelijne Muller, LLM, Prof. Virginia Dignum and Associate Prof. Aimee van Wynsberghe. Collectively, the founders have a broad expertise in AI: AI sciences, social impact, national and international policy, legal implications, and ethical impact.

CONTACT

ALLAI
Amsterdam Science Park 900 (LAB42)
1012 WX Amsterdam
The Netherlands

www.allai.nl
@ALLAI_EU
welkom@allai.nl