Kempelen Institute of Intelligent Technologies

Stance on the regulation of Generative Artificial Intelligence

October 2023

Abbreviations

AI means artificial intelligence.

AIA means proposal for Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

EU means European Union.

GDPR means regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

KInIT means Kempelen Institute of Intelligent Technologies with its registered seat in Bratislava, Slovakia.

Position of the Council means general approach of the Council of the European Union towards AIA adopted on 6 December 2022.

Position of the European Parliament means negotiation position of the European Parliament on the AI Act adopted on 14 June 2023.

Authors: Matúš Mesarčík, Natália Slosiarová, Juraj Podroužek, Mária Bieliková.

Table of contents

Abbreviations....2
Table of contents....3
Introduction....4
Executive summary....5
1. The scope of the regulation....7
2. Focus on short-term and current risks....8
3. Transparency....9
4. Privacy and data governance....11
5. Ex-ante auditability....12
6. Regulatory oversight....12
7. “Know-Your-Customer” checks....13
Acknowledgements....15
About us....16

Foreword

Whereas KInIT is an independent, research, non-profit institute focusing on artificial intelligence and related disciplines. KInIT's mission is to support scientific excellence and its transformation to responsible innovations by bridging the private and academic sectors, encouraging knowledge sharing, talent development and circulation, and advocating quality, ethics, and fairness including public policy advising.

Whereas artificial intelligence is at the core of our research.

Whereas our business and academic partners implement artificial intelligence systems in practice.

Whereas we appreciate the value of the public debate on societal impact of artificial intelligence in general.

Whereas we have already published our Stance on the Proposal for Artificial Intelligence Act in summer 2021.

Whereas we carefully observe the debate on benefits and risks of generative AI systems.

Whereas our researchers contributed to the Opinion on ethical issues of generative AI systems and large language models issued by the Commission for ethics and regulation of artificial intelligence established by the Ministry of Investments, Regional Development and Informatization of the Slovak Republic.

We are presenting our stance on the regulation of generative AI.

Executive summary

In this stance we are presenting our positions on the selected aspects of regulation of general purpose AIs, foundation models and generative AI systems as proposed by the positions of the European Parliament and Council towards Artificial Intelligence Act.

Our concerns primarily revolve around the correct definition of general purpose systems, foundation models and generative AIs. Further suggestions are made towards focusing on short-term risks, transparency, privacy and data governance, ex-ante auditability and regulatory oversight. We are also proposing know-your-customer checks.

In our view, the proposed definition by the Council of the EU is very broad and includes applications not specific to general-purpose AIs, e.g. translation.

In our opinion, the distinction between foundation models and general-purpose AI shall be more thoroughly explained. Furthermore, vague notions like broad scale or wide range of applications may be subject to restrictive interpretation from the providers thus escaping the scope of requirements.

We should focus on already existing risks posed by AI and generative AI in particular, including transparency, bias, privacy, human oversight or sustainability.

From the point of view of regulation, the limits of the deployment and use of generative AI systems must be clearly established, as well as the range of persons who can come into contact with them.We are of the opinion that users of generative AI systems should be informed that they are not interacting with a human, or that they are receiving output that has been machine generated. Providers or deployers should provide such information where the interaction with the system takes place, along with a warning that the generated output may also contain information that isn't true or verified. This warning shall be visible during the interaction with the generative AI.

We understand limited obligations towards generated content for purposes of freedom of speech. However, such exceptions if introduced shall be carefully balanced considering risks of generative AI.

We are of the opinion that providers of foundation models and generative AI systems shall document provenience of data sources used for training models together with information required by EU data protection law.

In our opinion, requirements for foundation models and generative AI shall be auditable with the aid of the third party, if such model or system is intended to be used in the high-risk area.

We believe that the European Commission is best suited to provide uniform and effective oversight over foundation models and generative AI.

We believe that general-purpose/ generative AI system/ foundation model providers shall conduct “Know-Your-Customer” checks to ensure that their models are used according to provided instructions, thus mitigating the risk of aiding any human rights abuse.

Introduction

In the spring 2021 the European Commission introduced the first EU comprehensive law on artificial intelligence - AI Act or AIA. In the following years, the Council of the EU and European Parliament published their positions on the original proposal. However, the underlying philosophy of the AIA remains.

AIA reflects a risk-based approach categorizing AI systems based on their impact on health, security and fundamental rights and freedoms. AIA bans several practices with AI systems (social credit scoring or unrestricted facial recognition technologies in public spaces used by law enforcement) but mainly focus on high-risk AI systems. Providers of high-risk AI systems are subject to a number of obligations with the aim to mitigate potential negative impacts on health, security and fundamental rights and freedoms. Providers of AI high-risk systems have to implement processes to mitigate potential biases in the AI systems, human oversight or robust data quality and governance practices. These obligations shall be documented during the process of conformity assessment. Conformity assessment is in majority of the cases conducted by providers of high-risk AI systems themselves. After successful completion, a high-risk AI system is registered in the EU database, receives declaration of conformity and may be introduced on the EU market. Robust oversight and post-market surveillance is foreseen by AIA on the national and EU level.

In the original proposal, generative AI systems were not explicitly recognized as high-risk AI systems and remained unregulated. However, with the public introduction of such systems, regulators aim to set forth rules for such systems. In this stance, we are presenting our views and suggestions towards regulation of generative AI systems.

1. The scope of the regulation

This stance primarily focuses on generative AI. However, our position takes into account broader categories of general-purpose AI and foundation models. It has to be mentioned that such systems and models were not governed in the original text of the AIA proposed by the European Commission in spring 2021. However, the introduction of generative AI tools for public[1] inevitably changed the course of the legislative process with the aim of the EU to regulate these systems.

The material scope of each regulation should be closely linked to the proper definition of basic notions. In case of generative AI, notions of general-purpose AI and foundation models shall be assessed as well. As we have discussed the original definition of AI systems elsewhere, we focus here on these three concepts without questioning the definition of AI provided in AIA.[2]

Position of the Council of the EU introduced the definition of general-purpose AI as *"an AI system that - irrespective of how it is placed on the market or put into service, including as open source software - 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."*[3]

In Council's position, general-purpose AI are further governed by obligations for risk management[4] that are hard to escape as the exception applies only when the provider has explicitly excluded all high-risk uses.[5] The notion does not take into account that generality of AI may relate to their ability, domain, tasks or output.[6]

> **In our view, the proposed definition by the Council of the EU is very broad and includes applications not specific to general-purpose AIs, e.g. translation.**

[1] Including ChatGPT, DALL-E or Midjourney.
[2] Mesarcik, M., Solarova, S., Podrouzek, J., Bielikova, M. Stance on The Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence – Artificial Intelligence Act. Kempelen Institute of Intelligent Technologies. September 2021. DOI: 10.31235/osf.io/yzfg8.
[3] Position of the Council of the EU on proposal for Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Article 3 sec. 1b.
[4] See Position of the Council, Articles 4a, 4b and 4c.
[5] Position of the Council of the EU, Article 4c.
[6] Gutierrez, Carlos Ignacio and Gutierrez, Carlos Ignacio and Aguirre, Anthony and Uuk, Risto and Boine, Claire and Franklin, Matija. A Proposal for a Definition of General Purpose Artificial Intelligence Systems (October 5, 2022). Available at SSRN: https://ssrn.com/abstract=4238951 or http://dx.doi.org/10.2139/ssrn.4238951.

Position of the European Parliament is more specific as it differentiates among general-purpose AI, foundation models and generative AI together with specific requirements. The definition of general-purpose AI is shorter than in the Position of the Council of the EU. General-purpose AI system is defined as an "*AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed*."[7]

The foundation model, according to the Position of the European Parliament, is an *"AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks."*[8] Generative AI is not defined in the article containing definitions but in the part specifying obligations on such systems as *"foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video."*[9]

> **In our opinion, the distinction between foundation models and general-purpose AI shall be more thoroughly explained. Furthermore, vague notions like broad scale or wide range of applications may be subject to restrictive interpretation from the providers thus escaping the scope of requirements.**

Additionally, it may be confusing to use the formulation "AI system model" as the word system seems redundant. According to our understanding, general-purpose AI systems are systems that might be used and utilized in different contexts and for different tasks. Foundation models might be fine-tuned for specific purposes. Additionally, generative AI systems can use foundation models for generating content. However, this is only one of the options in the relationship between foundation models and generative AI systems. Such distinctions are not clear from the definitions proposed by the EU Council or European Parliament.

2. Focus on short-term and current risks

In recent months, several initiatives were published calling for a temporary ban on the development of AI emphasizing long-term catastrophic risks.[10] We are also aware of some long-term risks posed by AI including the transformation of the labor market, or

[7] Position of the European Parliament on proposal for Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Article 3 sec. 1d.
[8] Position of the European Parliament, Article 3 sec. 1c.
[9] Position of the European Parliament, Article 28b sec. 4.
[10] Notably Future of AI, Pause Giant AI Experiments: An Open Letter, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

massive generation and spread of disinformation content. However, we are of the opinion that we need to focus on issues regarding AI systems including generative AI that are already developed, deployed and widely used.

This raises a question of whether this "long-termistic" mindset isn't counterproductive when assessing imminent AI risks and whether there are ways to mitigate risks without the need to fully ban AI development.

> **We should focus on already existing risks posed by AI and generative AI in particular, including transparency, bias, privacy, human oversight or sustainability.**

These risks shall be mitigated by compliance with principles and requirements for trustworthy AI.[11]

We welcome proposals from the European Parliament to include these principles explicitly in the legal framework.[12] The generality of the requirement being applicable to all AI systems shall be upheld.

3. Transparency

One of the most pressing issues concerning AI systems in general is anthropomorphization. This is especially evident in the case of generative AI as many users are not aware that they are interacting with content generated by machines or projecting human values on machines.

Generative AI is capable of producing outputs that are often unrecognizable from human artifacts.[13] Such effect may result in severe outcomes including a change in moral attitudes[14] or even suicide.[15] Among other things, this opens the door for massive generation and dissemination of disinformation content in the online space, manipulation of political attitudes and public opinion of the population, or undermining of trust in democratic institutions.

Generative AI systems should not create the false impression that their user is communicating with a human.

---

[11] European Commission High-Level Expert Group on AI. Ethics guidelines for trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[12] Position of the European Parliament, Article 4a.
[13] The Verge. The swagged-out pope is an AI fake — and an early glimpse of a new reality, https://www.theverge.com/2023/3/27/23657927/ai-pope-image-fake-midjourney-computer-generated-aesthetic.
[14] Krügel, S., Ostermaier, A. & Uhl, M. ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep 13, 4569 (2023). https://doi.org/10.1038/s41598-023-31341-0.
[15] Gintaras Radauskas. Belgian man commits suicide after talking to chatbot. Available at: https://cybernews.com/news/man-takes-own-life-chatbot/.

From the point of view of regulation, the limits of the deployment and use of generative AI systems must be clearly established, as well as the range of persons who can come into contact with them.

On the other hand, the providers and deployers of generative AI systems shall also ensure the protection of those persons for whom the use of generative AI systems represent a disproportionate risk of negative impact on their life, health and safety, especially children or other vulnerable groups. The identification of these groups as well as the method of their protection should be determined by legislation.

We are of the opinion that users of generative AI systems should be informed that they are not interacting with a human, or that they are receiving output that has been machine generated.

Providers or deployers should provide such information where the interaction with the system takes place, along with a warning that the generated output may also contain information that isn't true or verified.

This warning shall be visible during the interaction with the generative AI.

Therefore, we welcome proposed amendments to the AI Act by the European Parliament in the Article 52 dealing with transparency of AI systems. Transparency obligations enhanced towards text content is an important step especially relevant for fighting against disinformation online.

We understand limited obligations towards generated content for purposes of freedom of speech. However, such exceptions if introduced shall be carefully balanced considering risks of generative AI.

Furthermore, if copyrighted works are used for the purposes of training AI models, this shall be acknowledged by the provider of AI systems or other entities involved in training generative AI.

We welcome proposals from the European Parliament for providers of AI systems to publish sufficiently detailed summary of the use of training data protected under

copyright law.[16] The proposed mechanism is one of the feasible measures to promote transparency and inform copyright holders.

The opt-out model of data mining exceptions as provided by EU law[17] shall be upheld. However, we share concerns related to applicability and call for clearer formulation of the requirement.[18] Legislators shall carefully observe and reflect state of the art.

4. Privacy and data governance

Privacy and personal data protection are crucial requirements to respect considering generative AI especially in the phase of training. Several concerns have been already raised by national data protection authorities.[19]

The question of respecting privacy and data protection is not exclusive concerning potential data breaches, but also in case of training models.

Insufficient data transparency[20] shall be considered as one of the most pressing issues related to foundation models trained on a vast range of data from various sources including web scraping.

Data used for training may also include sensitive personal data that are subject to stricter requirements for processing.[21]

> **We are of the opinion that providers of foundation models and generative AI systems shall document provenience of data sources used for training models together with information required by EU data protection law.**

In general, we propose layered transparency of sources. Every foundation model shall have documented sources of data used for training. When foundation models are used and fine-tuned, the provider of the foundation model must adhere to the requirement of verification of sources and cooperate with the deployer or user. If the foundation model is publicly deployed, the public shall have information on used

[16] Position of the European Parliament, Article 28b (4) c).
[17] Notably Articles 3 and 4 of 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.
[18] QUINTAIS, R. Generative AI, Copyright and the AI Act. Available at: file:///Users/matusmesarcik/Zotero/storage/J6MS69GH/generative-ai-copyright-and-the-ai-act.html
[19] See e.g. Luca Bertuzzi. Italian data protection authority bans ChatGPT citing privacy violations. Available at: Italian data protection authority bans ChatGPT citing ...EURACTIV.comhttps://www.euractiv.com › news.
[20]Leigh Mc Gowran. OpenAI criticised for lack of transparency around GPT-4. Available at: https://www.siliconrepublic.com/machines/openai-gpt4-transparency-ai-concerns-stripe-chatgpt
[21] See Article 9 of the General Data Protection Regulation.

sources. Oversight bodies shall have full access to documents and datasets used for training.

5. Ex-ante auditability

The core obligation for high-risk AI systems according to AIA is to conduct conformity assessment - process of verification of requirements as set out by the regulation. This obligation also applies towards general-purpose AI as foreseen by the Position of the Council of the EU[22] and the Position of the European parliament for foundation models.[23]

However, as currently proposed, only a limited number of high-risk AI systems (biometric and biometrics-based systems) shall conduct conformity assessment with aid of the third party.

> **In our opinion, requirements for foundation models and generative AI shall be auditable with the aid of the third party, if such model or system is intended to be used in the high-risk area.**

Additionally, we welcome the obligation to conduct fundamental rights impact assessments on the side of deployers.[24] It is not clear from the wording of the AIA or respective positions if the obligation also applies to foundation models including generative AI in high-risk areas. We strongly encourage the legislator to include such requirements for these systems.

6. Regulatory oversight

General-purpose AI, foundation models and generative AI present different severity and types of risks. This is also evident from the approach taken by the European parliament governing foundation models and generative AI in a more elaborate manner.[25] We specifically want to draw attention to the requirement of their compliance with the obligation to identify, reduce and mitigate foreseeable risks to democracy and rule of law.[26]

[22] Position of the Council of the EU, Article 4b section 3.
[23] Position of the European Parliament, Article 28b section 2 letter f).
[24] Position of the European Parliament, Article 29a.
[25] Position of the European Parliament, Article 28b.
[26] Position of the European Parliament, Article 28b section 2 letter a).

The European Commission already possesses powers to enforce measures for the protection of rule law through the conditionality mechanism.[27] Furthermore, the European Commission is the sole oversight body for very large online platforms and very large search engines as stipulated by the Digital Services Act, a landmark piece of legislation governing transparency and accountability obligations for online media.[28]

> **We believe that the European Commission is best suited to provide uniform and effective oversight over foundation models and generative AI.**

Our recommendation is also stemming from the fact that generative AI represents a systemic risk[29] that may heavily influence online space, economy, mental wellbeing, environment and socio-technical landscape.

The oversight of such models and systems will therefore require a multidisciplinary approach concerning financial aspects, environmental aspects, governance of the digital market, cybersecurity, data governance, copyright issues or economic activities of providers of AI systems. We are of the opinion that the European Commission shall be the sole oversight body for such models and systems.

7. "Know-Your-Customer" checks

Due to the nature of general-purpose AI, generative AI and foundation models and their innumerable applications it is necessary to recognise that misuse can come from many sources within the downstream supply chain. Mere prohibition of misuse in the user manual is not enough to stop it and does not encourage accountability.

> **We believe that general-purpose/ generative AI system/ foundation model providers shall conduct "Know-Your-Customer" checks to ensure that their models are**

[27] Regulation (EU, Euratom) 2020/2092 of the European Parliament and of the Council of 16 December 2020 on a general regime of conditionality for the protection of the Union budget.
[28] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).
[29] Notion of systemic risk is already part of the EU law provisioned in macro-prudential oversight of the financial system legislation. "systemic risk means a risk of disruption in the financial system with the potential to have serious negative consequences for the real economy of the Union or of one or more of its Member States and for the functioning of the internal market. All types of financial intermediaries, markets and infrastructure may be potentially systemically important to some degree." Regulation (EU) No 1092/2010 of the European Parliament and of the Council of 24 November 2010 on European Union macro-prudential oversight of the financial system and establishing a European Systemic Risk Board, Article 2 c).

used according to provided instructions, thus mitigating the risk of aiding any human rights abuse.

This recommendation builds on a common best practice developed and applied in the past decades in the financial sector in order to protect against money laundering and financing criminal activities. A form of "Know-Your-Customer" (KYC) check was already proposed by Microsoft[30] and KYC checks in general are deemed as one of safety best practices also by OpenAI.[31]

There is also an obligation on a regulatory level similar to KYC checks. Article 28 of the GDPR stipulates that the controller shall conclude a due diligence check to ensure that the processor meets the requirements of the regulation and the rights of data subjects are protected during processing. Therefore, KYC checks may be built on an already existing practice in the EU data protection law.

This requirement shall, however, only be expected to apply to a limited number of customers and high-risk uses in order not to overly burden the providers.

[30] Microsoft. Governing AI: A Blueprint for the Future. Available at: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.
[31] OpenAI. Safety best practices. Available at: https://platform.openai.com/docs/guides/safety-best-practices.

Acknowledgements

The authors would like to thank the following people for their valuable comments and insights on the regulation and stance: **Matúš Pikuliak, Michal Gregor, Róbert Móro and Marián Šimko**.

About us

Kempelen Institute of Intelligent Technologies (KInIT) is an independent, non-profit research institute that conducts cutting-edge research on intelligent technologies, primarily focusing on artificial intelligence and its intersections with other disciplines. KInIT can be perceived as a catalyst for the Slovak innovation ecosystem in the area of intelligent technologies by conducting excellent research in artificial intelligence and its innovative applications. We are committed to connecting excellent science with innovative companies, their needs and experiences. Drawing inspiration from leading institutions in other countries, KInIT serves as a center of expertise that encourages companies to engage in research, strengthens their connections with the academic sector, and attracts talent to Slovakia. For more information about KInIT, please visit: https://kinit.sk/about-us/.