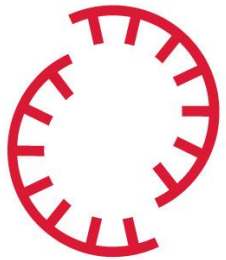
General Purpose, Generative or Foundation? How to regulate LLMs

filip.konopczynski@panoptykon.org



PANOPTYKON FOUNDATION

Panoptykon Foundation

Monitoring risks and legal abuses
In collaboration with other legal experts, we monitor legal solutions, both those in force and those proposed, including those at EU level. In co-operation with investigative journalists we are also trying to monitor the activity of public and private entities that use advance surveillance techniques in order to identify possible areas of abuse.

Legal interventions
We are trying to act as a public spokesman when human rights are threatened by the oppressive use of surveillance technologies. We prepare drafts of amendments to existing laws and legal opinions to new legislative proposals in the Polish and the EU Parliaments.

Research, public discussion, education
In co-operation with lawyers, sociologists, journalists and experts from other fields we analyse surveillance-related issues from various research perspectives. We publicize and comment on cases of abuse. Our media presence is constantly increasing.

Large Language Models & the law: what's the big deal?

Chat GPT became the fastest downloaded app in the history of the internet launching the new „AI frenzy”.

For many people, it was the first encounter with a „Touring-test” level piece of AI technology.

It also unleashed cultural and societal forces that were difficult to control.

What changed?

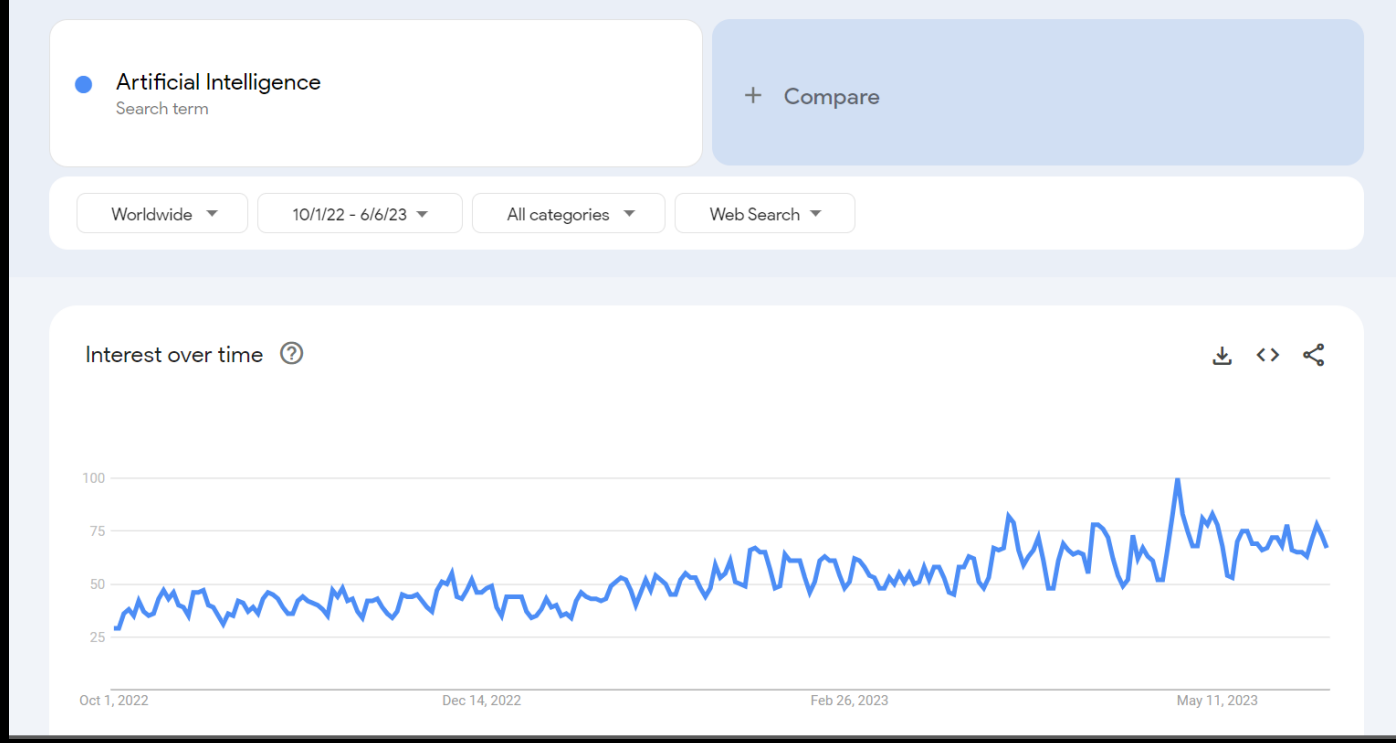
1. Moral panic and technological doomsaying („AI's Rorschach test"?)

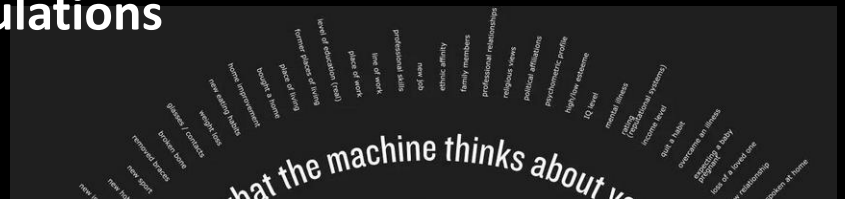
2. Finally raised the appeal of AI regulations for politicians

3. (Short) crisis in the Big Tech lobbying alliance?

4. It's no longer ok to say loudly „AI needs no regulations"

5. A new (although brief) window of opportunity for digital rights watchdogs and thinktanks to sway the discussion towards crucial issues with AI regulations





What remained the same?

1. Despite many new ideas, there is still no effective international AI cooperation

2. At the same time, the notion of „there is no law regulating AI” is a myth. Many laws: GDPR (privacy), DSA (platforms and services), DMA (market regulations), as well as cybersecurity regulations exist in the EU alone

3. Big Tech, despite the change of tone still engages in widespread lobbying campaigns in both the US and EU (main goals: not to treat LLMs as High Risk AI, establish regulations consistent with corporate compliance, push the obligations on deployers + lower the penalties)

4. Despite atempts to bring it to the global level, the battle for the future of AI regulations takes place in the EU (AI Act)

Detour: why so dystopian?

1. As soon as users started reporting „problems” with GPT, OpenAI’s CEO launched a media and lobbying offensive both in the US (DC, Congress), and the EU (Brussels, Madrid, Warsaw etc.).

2. While presenting exciting, and spectacular funtionalities of GPT4, Sam Altman often raised concerns about the „big” risks posed by AI:
   - jobs automation
   - possibility of sentient AI turning against humanity
   - „risk of extinction of humanity”

3. The same tone was spread by initiatives such as the „Open Letter” by Future of Life Institute (co-signed by many renown AI experts: Hinton, Bengio etc.), and more recently, Center for AI Safety. By setting the agenda in such a way, it’s easy shift the focus away from real life problems posed by modern AI systems and models.

4. ….In reality, the stakes are quite different – and more tangible. No matter what do we think about the prospect of AI apocalypse, it is certain that it can not be easily adressed. Unlike privacy, safety or intelectual property issues, that are currently violating existing laws and norms.

Legal risks and challenges (OECD, Panoptykon, 2023)

1. **Discrimination**, **exclusion** and **toxicity**: Harms that arise from the AI language model producing discriminatory and exclusionary **speech**

2. **Privacy I**: Are models trained on data obtained in a legitimate way (according to GDPR)?

3: **Privacy II**: Use of data by a LM can lead to data breaches through inadvertent leaking or inference of private information. AI language models can also facilitate both legitimate and illegitimate surveillance and censorship

4. Information hazards: Harms that arise from the AI language model leaking or inferring true sensitive information

5. **Disinformation** : AI language model producing false or misleading information can be employed by mal-intented actors (state, private, terrorist etc.)

6. **Malicious uses**: Harms that arise from actors using the AI language model to intentionally cause harm

7. Threat to intelectual property laws, as well as current media business model (monetization of content)

Societal and cultural risks (OECD, Panoptykon, 2023)

1. **Human-computer** interaction **harms**: Harms that arise from users overly trusting the AI language model or treating it as human-like

2. E**nvironmental harms**: Harms that arise from the AI language model's environmental or downstream economic impacts

3. **Assembly** line automation revolution for „white collars"? Fears of technological unemployment

4. **Financing** and other **barriers** for producers, deployers and users: will LLMs accelerate income inequalities (new tech controlled by a handful of companies)?

5. **Misinformation**: will LLMs blur the line between what's true and false even more?

6. **Mental health**: how chatbots using LLMs will affect mental health issues – particularly for children interacting with tchem?

AI Act: how LLMs changed the (political) game in Europe

1. AI Act negotiations started in 2021, when the European Commission presented the first draft. It raised interest mainly of Big Tech, legal experts, and some AI scholars and innovators

2. LLMs were not specifically tackled: moreover, the original definition of AI was ill-suited for them (included reference to „human-set objectives”)

3. Since people started talking about AI, it became easier to push politicians to fight for basic rights of people affected by AI: right to explanation, to lodge complaints (AI office, national authorities, courts), to notification, as well as introduce the *fundamental rights impact assessment* mechanism for H-R systems (art. 29a).

….But what about LLMs?

LLMs in AI Act: different ideas, same goals

At different stages of the AIA negotiation proces different ideas were being floated around, the main ones being:

1. No mention in the original draft by the EC.

2. Introduction of *General Purpose AI* in the Council's proposal:

Article 3(3b): „**'general purpose AI system'** means an AI system ((…)including open source software) 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"

- Requirements: treated close to High Risk Systems (art. 16-16j; 25,48, 61);
- BUT: if provider included „instructions excluding H-R uses", no mandatory compliance – unless informed about deployers' abuses, which obliges them to mitigate the risks (even turn off access to the service?)

LLMs & AI Act: current state of affairs (before the final vote)

That did not stand, and most likely (according to Europe Corporate Observatory's report) due to advocacy efforts by digital companies resulted in an introduction of a new term into the text adopted by the IMCO/LIBE committeees.

1. Definition of Foundation Models (Art. 3):

(1c) '**foundation model'** means an AI 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;

(1d) 'general purpose AI system' means 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.

2. Generative AI – despite widespread use, no such definitions in the AI Act (specifically regulated only in art. 52)

3. General Purpose AI – remained in the text, but lost its original purpose

Obligations of **providers** of FM (AI Act, EP, 06.23)

Article 28b

2. Provider of FM has to ensure that it is compliant with the requirements:

(a) demonstrate (through appropriate design, testing, 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;

b) process and incorporate only datasets that are subject to appropriate data governance measures for foundation models (sustainability, bias);

c) achieve throughout its lifecycle appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity (including involvement of independent experts);

d) Sustainable energy use, resource use and waste, as well as to increase energy efficiency;

e) extensive technical documentation and intelligible instructions for use in order to enable the downstream providers to comply with their obligations;

f) quality management compliance system;

g) register that foundation model in the EU database (art. 60)

3. keep the technical documentation for up to 10 (!) years

4. Generative FM Models:

- transparency obligations (art. 52(1);
- adequate safeguards against the generation of content in breach of EU law;
- detailed summary of the use of training data protected under copyright law.

5. Also applicable: General Principles (art. 4a): ‘human agency and oversight’; ‘technical robustness and safety’; ‘privacy and data governance’; ‘transparency’; ‘diversity, non-discrimination and fairness’; ‘social and environmental well-being’.

Obligations of **deployers** of AI systems based on FM (AI Act, EP 06.23)

1. Obligations under ANNEX VIII:
- applicable ONLY to certain High Risk AI deployers (public institutions, VLOPs [->DSA], also voluntarily for all)
- Basic information on FM: Name, address, contact details of the provider; description: of data sources that the model was built/trained on, capabilities and limitations, training resources incl. Computer power; performance benchmark; results of testing; url.

2. Transparency (art. 52):
- obligation to inform users about the use of chatbots and generative AI (photos, audio, video).-

3. And, last but not least: fines: „Non-compliance of AI system or foundation model with any requirements or obligations (other than those Articles 5, and 10 and 13) shall be subject to administrative fines of up to 10 000 000 EUR or, if the offender is a company, up to 2% of its total worldwide annual turnover for the preceding financial year".

4. In the case of HR AI systems (Article 28(a)) - provision against unfair contractual terms unilaterally imposed on SME or startups.

Summary: Foundation Models won't be treated as high risk systems, and most of the responsibilities are placed on the deployers.

What's next?

Despite achieving many of their goals, technological companies are not done on the lobbying front. Here's what to expect in the international context:

1. US-EU cooperation. The talks launched by the EP and the EC will continue via ministerial-level meeting of the Trade and Technology Council (TTC).

2. Global „AI Pact" – free-to-join initiative by the European Commission dedicated to AI companies to join before AI Act is implemented.

3. OECD and the Council of Europe are also working on international standards (insignificant, due to the lack of enforcement).

4. Sam Altman's United Nations AI Agency? Nice idea, but not realistic (no political chances of an effective, UN-level treaty).

5. Still: the only sheriff in town is the AI Act: if passed by the end of 2023, it becomes the law (2025?).

AI Act: Takeways

1. Thanks to hype in LLMs NGOs like Panoptykon and EDRi network had a chance to influence the debate over AI regulations in the EU. Unfortunately, SMEs, researchers and small innovators were not at the table at the final stages.

2. LLMs took regulators by surprise. In the EU, due to the nature of the legislative proces watchdogs, NGOs, SMEs/ innovators nor scientists had no real say - last amendments in the IMCO/LIBE were (most likely) created by and for large companies developing their own LLMs.

3. If passed, AI Act will set an EU and potentially global standard for LLMs like GPT, Bard or LaMDa. It puts certain requirements on providers, but mainly shifts the responsibility to deployers – people, institutions and companies who want to utilize FMs. The requirements on providers are not drastic, but require both technical capacity, and legal & compliance suport.

4. Will it „stifle innovation" for start-ups? Rather not, but If compliance obligations turn out to be too cumbersome, it should be adressed by providing sufficient external funding dedicated to fullfilling new regulatory obligations related to fundamental rights of people.

AI Act - civil society (EDRi) demands in the last stages of the negotiations

1. **Empower people affected by AI systems:**
   - The right to seek information when affected by AI-assisted decisions and outcomes;
   - A right for people affected to lodge a complaint with a national authority, if their rights have been violated by the use of an AI system;
   - A right to representation of natural persons and the right for public interest organisations to lodge standalone complaints with a national supervisory authority;
   - rights to effective remedies for the infringement of rights;
   - Access and ability standards for everybody using AI systems.

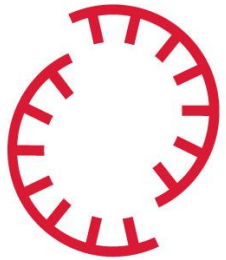
2. **Ensure accountability and transparency for the use of AI:**

   - An obligation on deployers to conduct and publish a fundamental rights impact assessment before each deployment of a high-risk AI system
   -Require all deployers of all high-risk AI systems to register the use in the European AI database before deployment;
   - The classification process for high-risk AI systems prioritizes legal certainty and provides no loophole for providers to circumvent legal scrutiny;
   - EU-based AI providers whose systems impact people outside of the EU are subject to the same requirements as those inside the EU.

3. **Prohibit AI systems that pose an unacceptable risk for fundamental rights:**

   - real-time and post remote biometric identification in publicly accessible spaces, by all actors, without exception;
   - all forms of predictive and profiling systems in law enforcement and criminal justice (location / place-based and person-based);
   - individual risk assessments and profiles based on personal and sensitive data, and predictive analytic systems when used to interdict, curtail and prevent migration; biometric categorisation systems that categorise natural persons according to sensitive or protected attributes as well as the use of any biometric categorisation and automated behavioural detection systems in publicly accessible spaces;
   - emotion recognition systems to infer people's emotions and mental states from physical, physiological, behavioural, as well as biometric data.

Thank you



PANOPTYKON FOUNDATION

Panoptykon Foundation
www.panoptykon.org
fundacja@panoptykon.org
filip.konopczynski@panoptykon.org

Selected sources

- AI language models: Technological, socio-economic and policy considerations, OECD 2023
- Governing AI: A Blueprint for the Future, Microsoft 2023
- Generating Harms, GenerativeAI'sImpact & PathsForward 2023
- Chat GPT – about time to regulate it, Panoptykon 2023
- What you need to know about generative AI and human rights
- The Lobbying Ghost in the Machine, Corporate Europe Observatory, 2023
- AI Act proposal – EP (consolidated version before the 14.06 vote in the European Parliament), 2023