Flip or Flop?

The Pacing Problems of Systemic Risk GPAI

Joint paper with Jockum Hildén, Research Institute for Sustainable AI (RISAI), and Mannheimer Swartling

Stefan Larsson

Associate Professor in Technology and Social Change,

Lawyer (LLM)
PhD in Sociology of Law
PhD in Spatial Planning

Department of Technology and Society, Faculty of Engineering, Lund University, Sweden.

The XXXV Nordic Conference in Law and IT, Artificial Intelligence and Legal Methods – Navigating the New Frontier, 5-6 November 2024.

WASP—HS

FUTURE CHALLENGES IN THE NORDICS

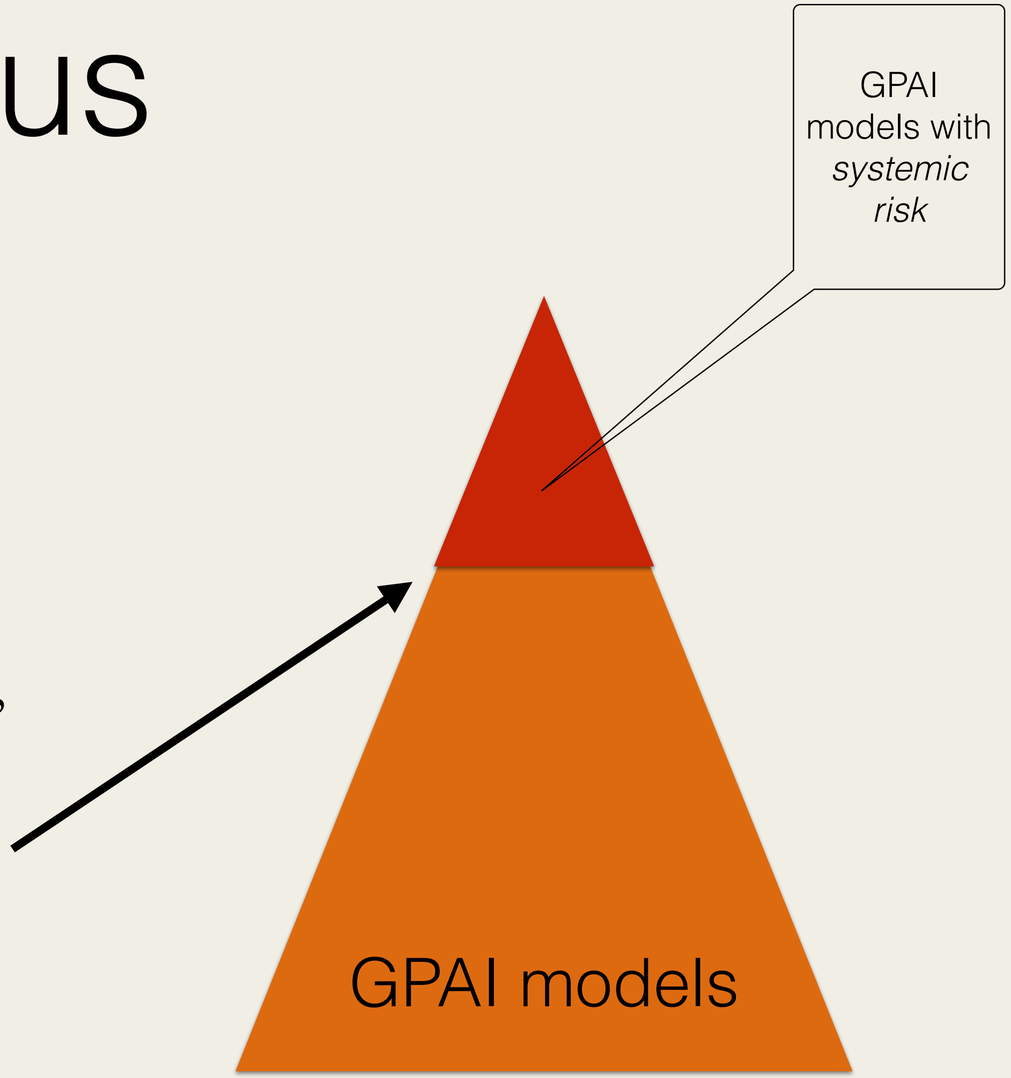
LUND UNIVERSITY

Main focus

1. **The leap** from *computation* (measured in FLOPs) to *systemic risk* that can have

"negative effects on public health, safety, public security, fundamental rights, or the society as a whole"

2. **The flexible**/insecure/discretionary regulation of systemic risk designation in GPAI.



My research context

- **The Automated Administration: Governance of ADM in the public sector** (PI: Stefan, 2022-2026), funder: SLS, Future Challenges in the Nordics

- **Exploring the risk governance mechanisms under the forthcoming EU AI Act** (PI: James White, 2024-2026), funder: Swedish Research Council

- **Vulnerability in the Automated State** (PI: Sofia Ranchordás, Jannice Käll, 2023-2027), funder: WASP-HS

- **AI Transparency and Consumer Trust** (PI: Stefan, 2019-2024), funder: WASP-HS

Digital Society (2024) 3:41
https://doi.org/10.1007/s44206-024-00129-8

Enforcement Design Patterns in EU Law: An Analysis of the AI Act

Kasia Söderlund[1] · Stefan Larsson[1]

Received: 1 November 2023 / Accepted: 1 July 2024
© The Author(s) 2024

Abstract
In recent decades, the enforcement of European Union (EU) law has transitioned from being primarily the responsibility of Member States to becoming an increas-

INTERNET POLICY REVIEW
Journal on internet regulation
Volume 9 | Issue 2

Transparency in artificial intelligence

Stefan Larsson
Department of Technology and Society, Lund University, Sweden, stefan.larsson@lantm.lth.se

Fredrik Heintz
Department of Computer Science (IDA), Linköping University, Sweden

International Journal of Social Robotics
https://doi.org/10.1007/s12369-023-01042-9

Towards a Socio-Legal Robotics: A Theoretical Framework on Norms and Adaptive Technologies

Stefan Larsson[1] · Mia Liinason[2] · Laetitia Tanqueray[1] · Ginevra Castellano[3]

Accepted: 3 August 2023
© The Author(s) 2023

Abstract
While recent progress has been made in several fields of data-intense AI-research, many applications have been shown to be prone to unintendedly reproduce social biases, sexism and stereotyping, including but not exclusive to gender. As more of these design-based, algorithmic or machine learning methodologies, here called *adaptive technologies*, become embedded in robotics, we see a need for a developed understanding of what role social norms play in social robotics, particularly with regards to fairness. To this end, we (i) we propose a framework for a *socio-legal robotics*, primarily drawn from Sociology of Law and Gender Studies. This is then (ii) related to already established notions of acceptability and personalisation in social robotics, here with a particular focus on (iii) the interplay between adaptive technologies and social norms. In theorising this interplay for social robotics, we look not only to current statuses of social robots, but draw from identified AI-methods that

Article

MJ

User accounts: How technological concepts permeate public law through the EU's AI Act

Maastricht Journal of European and Comparative Law
1–21
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1023263X241248469
maastrichtjournal.sagepub.com
Sage

Ida Koivisto*, Riikka Koulu**, and Stefan Larsson***

Between Regulatory Fixity and Flexibility in the EU AI Act

Abstract
The EU AI Act aims to regulate artificial intelligence (AI) in a way that balances innovation and protection from harms, but faces the challenge of keeping pace with the fast and dynamic development of AI. As a basis to address recent changes in the regulatory AI landscape, this paper examines the tension between fixity and flexibility when regulating emerging technologies, drawing on literature on the so-called pacing problem, contrasted by sociolegal theory on the importance of predictability and legal certainty. Specifically, it analyses how the EU AI Act, under the aim of being "future-proof" according to relatively newfound EU terminology, employs various mechanisms of flexibility, such as i) voluntary measures and codes-of-conduct, ii) delegated and implementing acts, and iii) harmonised standards to cope with the uncertainty and complexity of AI – potentially at the expense of predictability. The study includes aspects of how the AI Act addresses the emergence of *general-purpose AI* and *generative AI*, to illustrate challenges associated with regulating rapidly developing technologies. In conclusion, the paper argues that while flexibility is unavoidable when drafting law explicitly targeting such a swiftly moving and conceptually blurry field and concept as AI, it also entails trade-offs such as reduced legal predictability, which is concerning since predictability is essential for ensuring trust and legal certainty in the regulatory framework around this set of technologies, as well as a shift in powers, to the Commission and standardisation organisations.

Keywords: *AI Act, the pacing problem, legal certainty, legal flexibility, general purpose AI, delegated acts, harmonised standards*

1. Introduction and Purpose of the Study

Predictability is, on the one hand, since long a commonly emphasised essential feature of law. To uphold the principle of legality and legal certainty, legal scholars such as Alexander Peczenik state, predictability is one of the basic democratic values in and a state governed by legislation.[1] Legal sociologists like Vilhelm Aubert have expressed that the ability to secure expectations is one of the five main tasks of law,[2] and principles of proper law-making emphasise conceptual clarity and low frequency in changes.[3] In brief, the *fixity of norms*, to this foundational legal school of thought, is key.

On the other hand, in Europe and the Western world, there is an ongoing and more recent discourse regarding the dynamic between the demands for regulation and the opportunities and risks of

Original Research Article

DIGITAL HEALTH

Engaging with artificial intelligence in mammography screening: Swedish breast radiologists' views on trust, information and expertise

DIGITAL HEALTH
Volume 10: 1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241287958
journals.sagepub.com/home/dhj
Sage

Charlotte Högberg[1], Stefan Larsson[1] and Kristina Lång[2,3]

frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 09 October 2024
DOI 10.3389/frobt.2024.1456613

Check for updates

OPEN ACCESS

EDITED BY
Silvia Rossi,
University of Naples Federico II, Italy

REVIEWED BY
Maria Chiara Caschera,
National Research Council (CNR), Italy
Dana Vrajitoru,
Indiana University South Bend, United States

*CORRESPONDENCE
Kashyap Haresamudram,
kashyap.haresamudram@lth.lu.se

RECEIVED 28 June 2024
ACCEPTED 20 September 2024
PUBLISHED 09 October 2024

CITATION
Haresamudram K, Torre I, Behling M, Wagner C and Larsson S (2024) Talking body: the effect of body and voice anthropomorphism on perception of social agents.
Front. Robot. AI 11:1456613.
doi: 10.3389/frobt.2024.1456613

COPYRIGHT
© 2024 Haresamudram, Torre, Behling, Wagner and Larsson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with

Talking body: the effect of body and voice anthropomorphism on perception of social agents

Kashyap Haresamudram[1*], Ilaria Torre[2], Magnus Behling[3], Christoph Wagner[3] and Stefan Larsson[1]

[1]Department of Technology and Society, Lund University, Lund, Sweden, [2]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden, [3]Department of Economics, Lund University, Lund, Sweden

Introduction: In human-agent interaction, trust is often measured using human-trust constructs such as competence, benevolence, and integrity, however, it is unclear whether technology-trust constructs such as functionality, helpfulness, and reliability are more suitable. There is also evidence that perception of "humanness" measured through anthropomorphism varies based on the characteristics of the agent, but dimensions of anthropomorphism are not highlighted in empirical studies.

Methods: In order to study how different embodiments and qualities of speech of agents influence type of trust and dimensions of anthropomorphism in perception of the agent, we conducted an experiment using two agent "bodies", a speaker and robot, employing four levels of "humanness of voice", and measured perception of the agent using human-trust, technology-trust, and Godspeed series questionnaires.

Results: We found that the agents elicit both human and technology

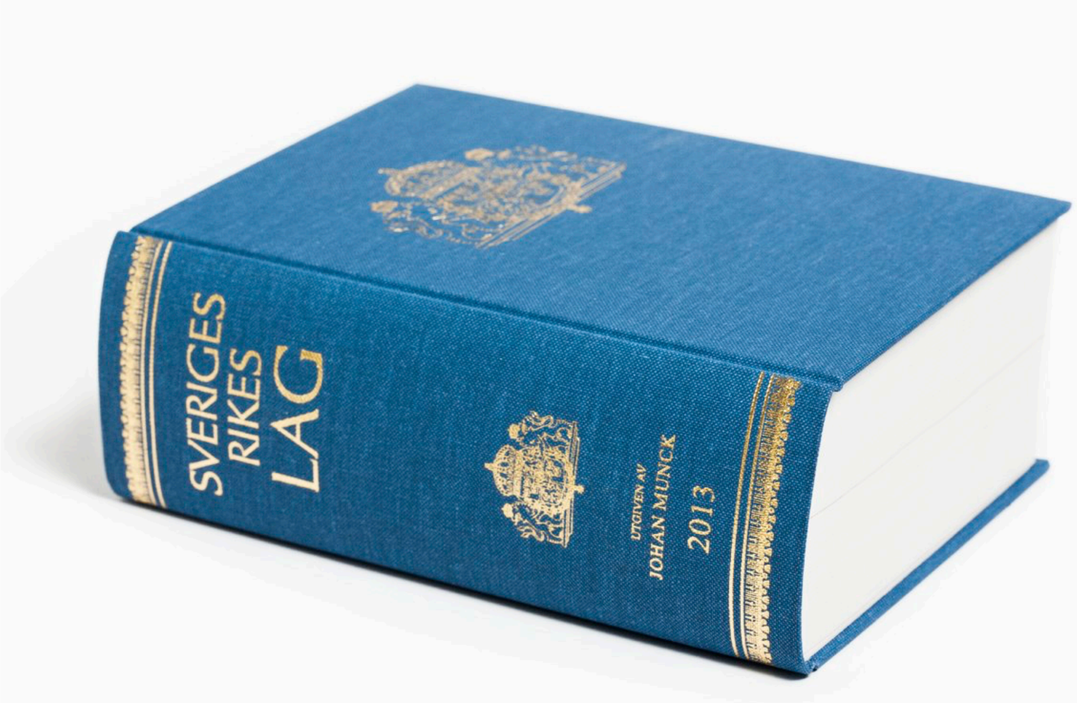
Tentative theoretical approach

Legal flexibility (the pacing problem)

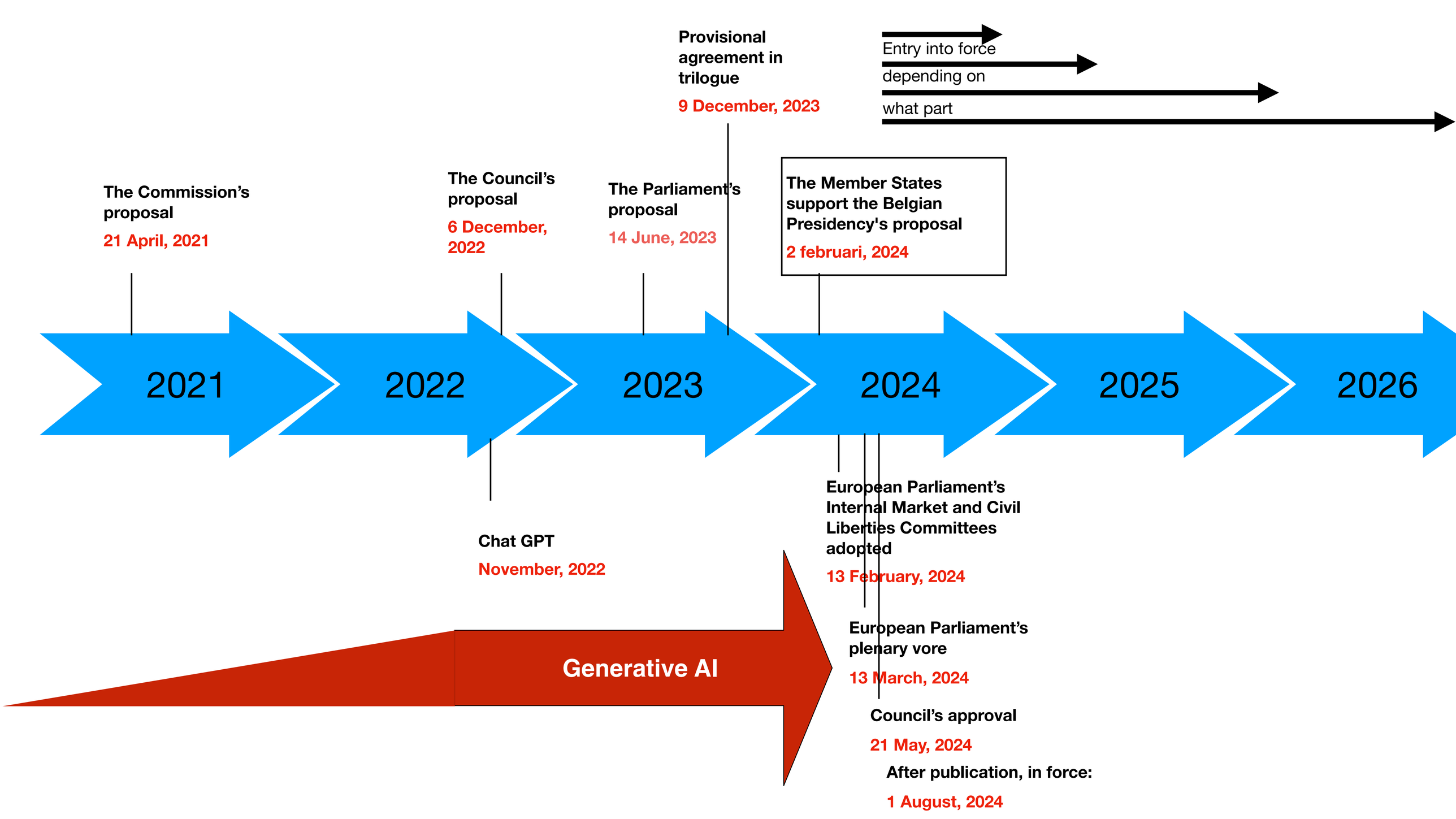
- **Soft** governance, soft law; experimental regulation and regulatory sandboxes (Ranchordas, 2021)
- “Future-proof”, **anticipatory** governance (Ranchordas & van ’t Schip, 2020; Mandel, 2020)
- ”**Co-regulation**” in terms of codes of practice, standardisation, certification
- Routines for **continuous technology monitoring** and risk assessments

Legal fixity (predictability)

- Need for legal **predictability** (Aubert; Peczenik etc.)
- Principles of proper lawmaking: **clarity in terminology, no frequent changes**, etc. (cf. Popelier, 2000)
- The **presence of the normative past**: Displacing decisions across time (Super, 2011)
- **In essence:** Certain level of temporal and conceptual fixity



The AI Act process


Provisional agreement in trilogue
9 December, 2023
Entry into force
depending on
what part
The Commission's proposal
21 April, 2021
The Council's proposal
6 December, 2022
The Parliament's proposal
14 June, 2023
The Member States support the Belgian Presidency's proposal
2 februari, 2024
2021
2022
2023
2024
2025
2026
Chat GPT
November, 2022
European Parliament's Internal Market and Civil Liberties Committees adopted
13 February, 2024
Generative AI
European Parliament's plenary vore
13 March, 2024
Council's approval
21 May, 2024
After publication, in force:
1 August, 2024

Signs of pace: Generative AI

- **Text:** GPT-3/-4/ChatGPT etc.
- **Image:** DALLE2/3; Midjourney; Stable Diffusion etc. (note how much that has happened since “speaking Mona Lisa”, 2019)
- **Sound:** VALLE etc. (note Darth Vader / James Earl Jones)
- **Video or “multimodal”:** within shortly (about now — note Sora in February).

Btw, another sign of pace?

Artikel 51

Klassificering av AI-modeller för allmänna ändamål som AI-modeller för allmänna ändamål med systemrisk

1. En AI-modell för allmänna ändamål ska klassificeras som en AI-modell för allmänna ändamål med systemrisk om den uppfyller något av följande villkor:

a) Den har kapacitet med hög påverkansgrad som utvärderats på grundval av lämpliga tekniska verktyg och metoder, inbegripet indikatorer och riktmärken.

b) Den har, baserat på ett beslut av kommissionen, på eget initiativ eller efter en kvalificerad varning från den vetenskapliga panelen, kapacitet eller inverkan som motsvarar den som avses i led a med beaktande av kriterierna i bilaga XIII.

2. En AI-modell för allmänna ändamål ska förutsättas ha kapacitet med hög påverkansgrad enligt punkt 1 a om den sammanlagda beräkningsmängd som används för dess träning mätt i flyttalsberäkningar är större än 1025.

Svenska

2. En AI-modell för allmänna ändamål ska förutsättas ha kapacitet med hög påverkansgrad enligt punkt 1 a om den sammanlagda beräkningsmängd som används för dess träning mätt i flyttalsberäkningar är större än 1025.

Engelska

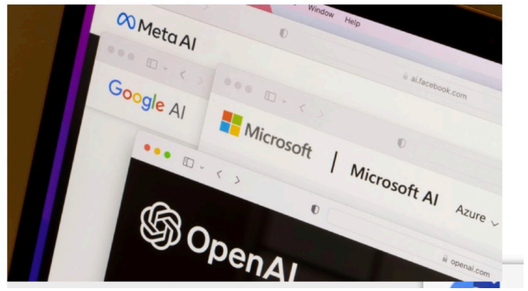
2. A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}.

EURACTIV

AI Act: EU countries headed to tiered approach on foundation models amid broader compromise

By Luca Bertuzzi | Euractiv.com **Est. 7min**

17 okt. 2023 (updated: 18 okt. 2023)



17 October

EURACTIV

EU's AI Act negotiations hit the brakes over foundation models

By Luca Bertuzzi | Euractiv.com **Est. 6min**

10 nov. 2023 (updated: 15 nov. 2023)

Content-Type: News

[Philippe STIRNWEISS/European Parliament]

10 November

EURACTIV

AI Act: Spanish presidency makes last mediation attempt on foundation models

By Luca Bertuzzi | Euractiv.com **Est. 6min**

29 nov. 2023 (updated: 1 dec. 2023)

Content-Type: News

Carme Artigas (left) is Spain's State Secretary for

29 November

Trilogue 2023 timeline

AI Act in Europe (9 Dec 2023)

The GPAI regulation

‘AI system’

“…a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” (Art 3 (1))

‘general-purpose AI model’

“…an AI model, including where such an AI model is trained with a **large amount of data** using **self-supervision at scale**, that displays **significant generality** and is capable of **competently performing a wide range of distinct tasks** regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, **except** AI models that are used for research, development or prototyping activities before they are placed on the market” (Art 3 (63))

Recital 99

“**Large generative AI models** are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.”

‘systemic risk’

“… a risk that is specific to the **high-impact capabilities** of general-purpose AI models, **having a significant impact** on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain” (Art 3 (65))

‘high-impact capabilities’

“… capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models” (Art 3 (64))

Recital 110

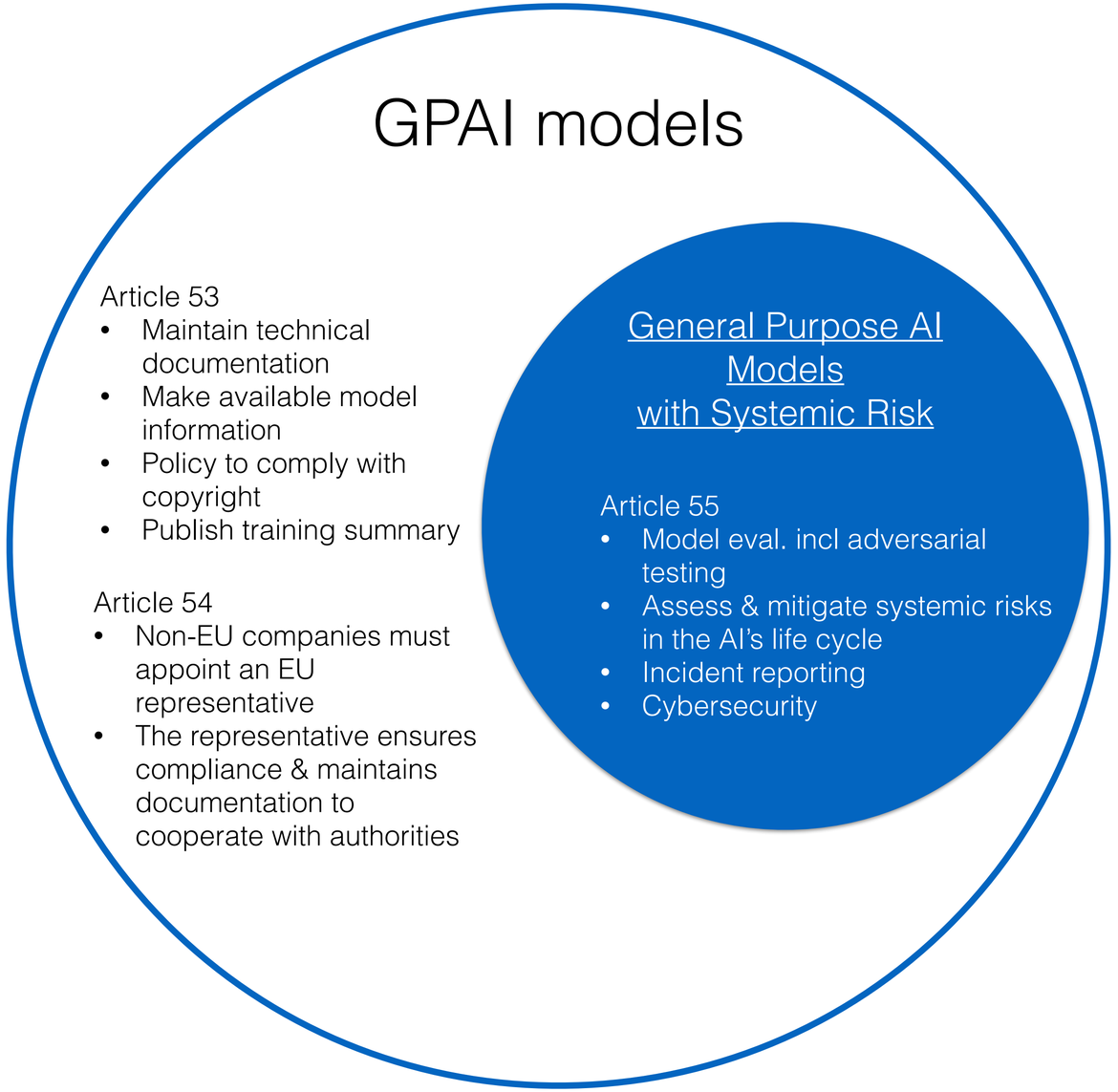
- …include, **“negative effects”** in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; …on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content.

- …increase with model **capabilities** and model **reach**

- …**relating to alignment with human intent**; chemical, biological, radiological, and nuclear risks

- ..the facilitation of disinformation or harming privacy with **threats to democratic values and human rights**; risk that a particular event could lead to a **chain reaction** with considerable negative effects that could affect up to an **entire city**, an **entire domain activity** or **an entire community**

Obligations


GPAI models
Article 53
• Maintain technical documentation
• Make available model information
• Policy to comply with copyright
• Publish training summary
Article 54
• Non-EU companies must appoint an EU representative
• The representative ensures compliance & maintains documentation to cooperate with authorities
General Purpose AI Models with Systemic Risk
Article 55
• Model eval. incl adversarial testing
• Assess & mitigate systemic risks in the AI's life cycle
• Incident reporting
• Cybersecurity

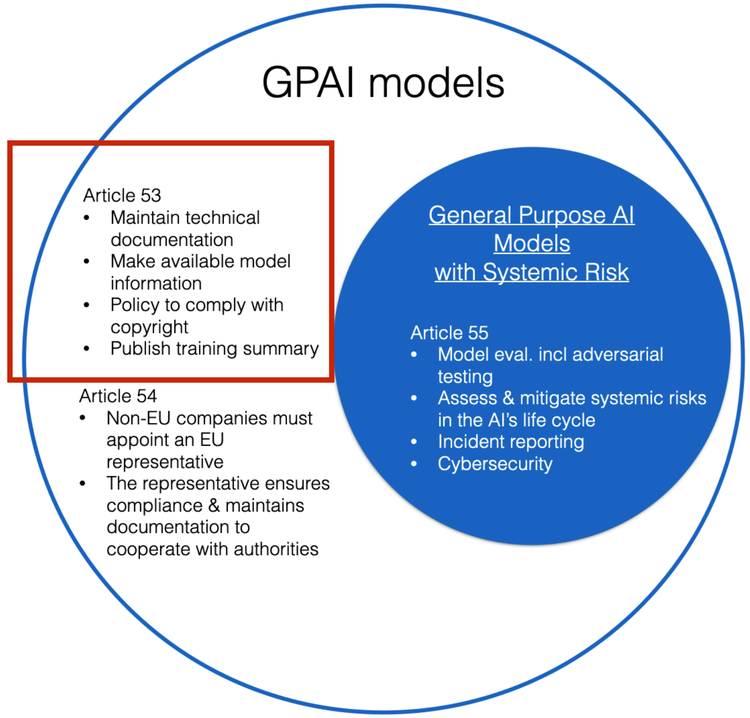
Obligations for Providers of General-Purpose AI Models

- **Draw up and keep up-to-date the technical documentation**…for the purpose of providing it, upon request, to the AI Office and the national competent authorities (Art 53(1a))

- **Make available information and documentation to providers** of AI systems who intend to integrate the general-purpose AI model into their AI systems (Art 53(1b))

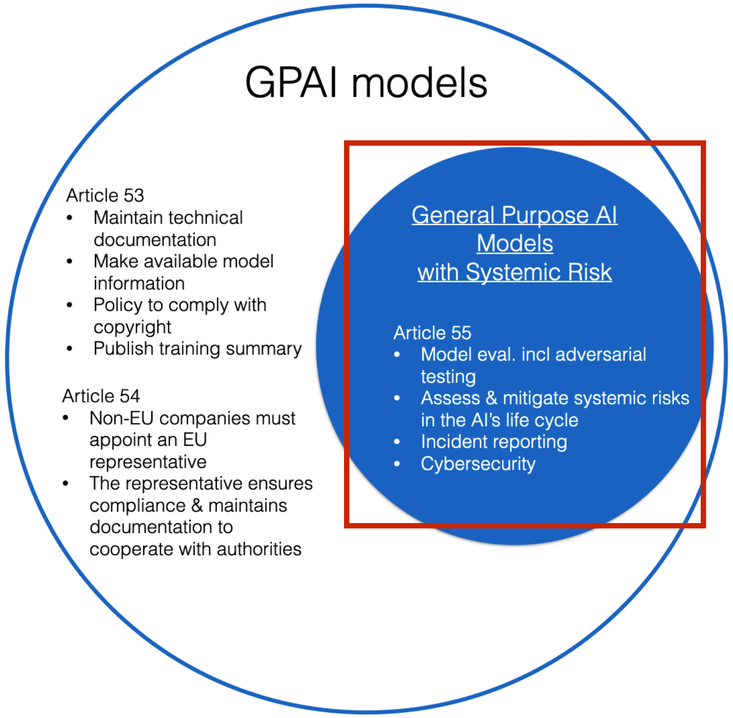
- **Put in place a policy to comply with Union law on copyright** and related rights (Art 53(1c))

- Draw up and make publicly available a **sufficiently detailed summary about the content used for training** (Art 53(1d))



Obligations of Providers of General-Purpose AI Models with Systemic Risk

- In addition to above…perform **model evaluation** …with a view **to identifying and mitigating** systemic risks (Art 55(1a))
- ..assess and mitigate possible systemic risks at Union level (Art 55(1b))
- Keep track of, document, and **report**, without undue delay, to the AI Office and, as appropriate, to national competent authorities… (Art 55(1c))
- Ensure an adequate level of **cybersecurity** (Art 55(1d))
- *And*: may rely on **codes of practice**…until a **harmonised standard** is published (Art 55(2))



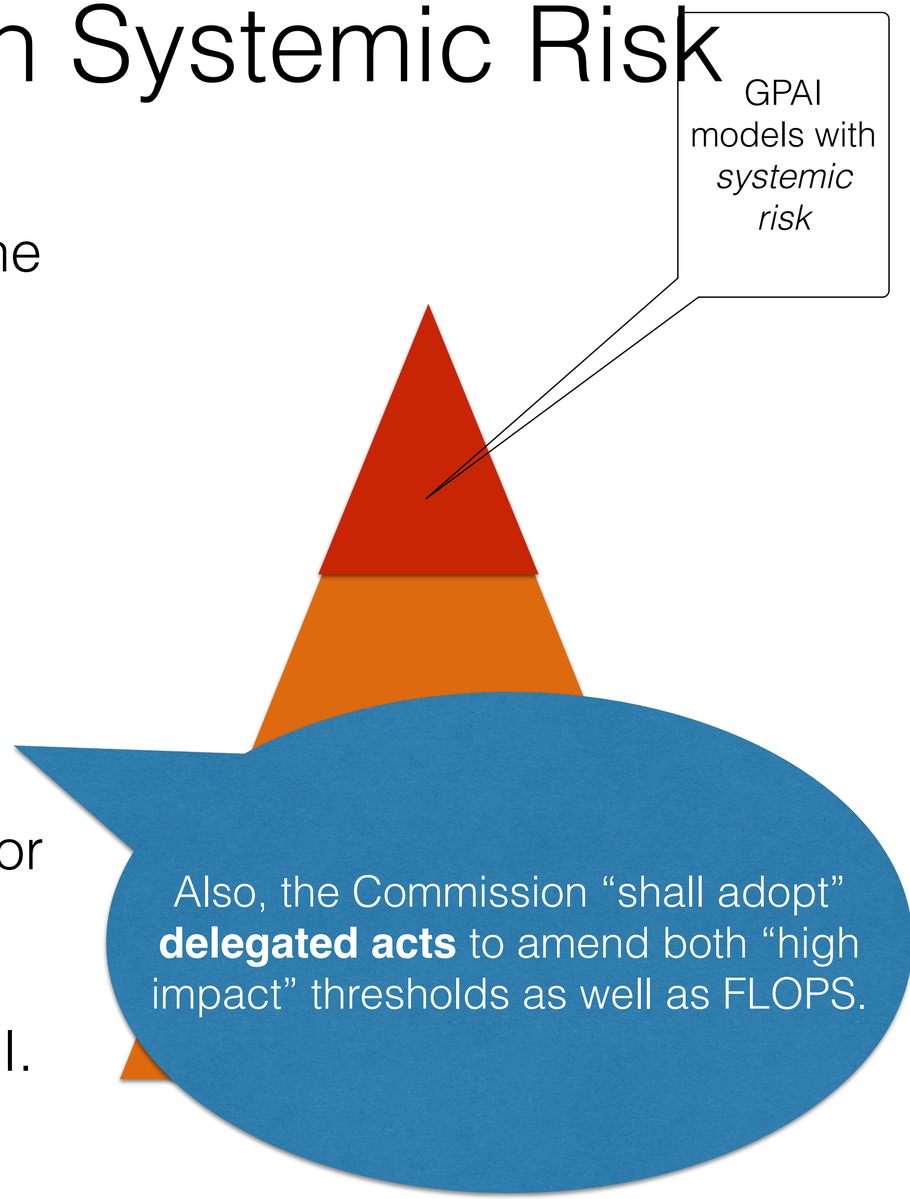
How is the line drawn?

Classification GP AI Models with Systemic Risk

- ***Either*** **if it has *high impact capabilities*** evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks (Art 51(1a))

  - ➡ **This is presumed** when *the cumulative amount of computation* used for its training measured in floating point operations is greater than 10(^25) — hence FLOPs (Art 51(2)).

- ***Or*** **based on a decision of the Commission,** ex officio or following a qualified alert from the scientific panel, it has capabilities or an impact equivalent to those set out in point (a) having regard to the criteria set out in Annex XIII.



Annex XIII

- The Commission should take these into account:

(a) the **number of parameters** of the model; (b) the **quality or size** of the data set, for example measured through tokens;

(b) the **amount of computation** used for training the model, measured in FLOPS or indicated by a **combination of other variables** such as estimated cost of training, estimated time required for the training, or estimated energy consumption for the training;

(c) the input and output modalities of the model…

(d) the benchmarks and evaluations of capabilities of the model, including considering the **number of tasks** without additional training, **adaptability to learn** new, distinct tasks, its **level of autonomy** and **scalability**, the tools it has access to;

(e) whether it has a high impact on the internal market due to its reach, which shall be presumed when it has been made available to at least 10 000 **registered business users** established in the Union;

(f) the number of **registered end-users**.

Who?

“Why is 10^25 FLOPs an appropriate threshold for GPAI with systemic risks?” (12 Dec 2023)

- “This threshold captures the currently most advanced GPAI models, namely OpenAI's GPT-4 and likely Google DeepMind's Gemini.”
- ”The capabilities of the models above this threshold are not yet well enough understood. They could pose systemic risks, and therefore it is reasonable to subject their providers to the additional set of obligations.”

European Commission - Questions and answers

Artificial Intelligence – Questions and Answers*

Brussels, 12 December 2023

Why do we need to regulate the use of Artificial Intelligence?

The potential benefits of Artificial Intelligence (AI) for our societies are manifold from improved medical care to better education. Faced with the rapid technological development of AI, the EU decided to act as one to harness these opportunities.

The EU AI Act is the world's first comprehensive AI law. It aims to address risks to health, safety and fundamental rights. The regulation also protects democracy, rule of law and the environment.

While most AI systems will pose low to no risk, certain AI systems create risks that need to be addressed to avoid undesirable outcomes.

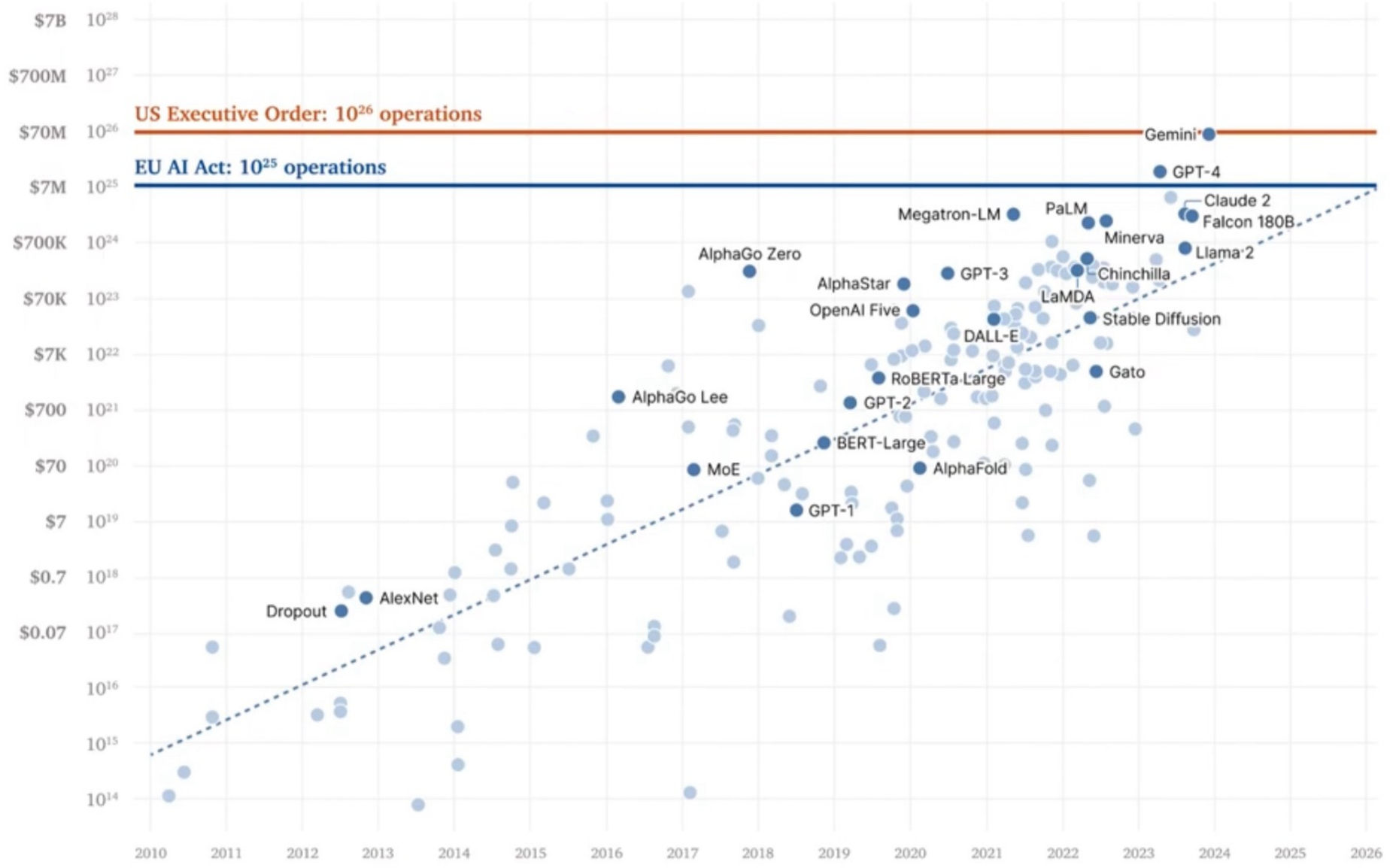
For example, the opacity of many algorithms may create uncertainty and hamper the effective enforcement of the existing legislation on safety and fundamental rights. Responding to these challenges, legislative action was needed to ensure a well-functioning internal market for AI systems where both benefits and risks are adequately addressed.

This includes applications such as biometric identification systems or AI decisions touching on important personal interests, such as in the areas of recruitment, education, healthcare, or law enforcement.

Recent advancements in AI gave rise to ever more powerful Generative AI. So-called “general-purpose AI models” that are being integrated in numerous AI systems are becoming too important for the economy and society not to be regulated. In light of potential systemic risks, the EU puts in place effective rules and oversight.

Compute Thresholds as Specified in the US Executive Order 14110 and EU AI Act

Estimated compute cost and total training compute used to train notable AI models, measured in total FLOP (floating-point operations) | Logarithmic



"Computing Power and the Governance of Artificial Intelligence"
Sastry, Heim, Belfield, Anderljung, Brundage, Hazell, O'Keefe, Hadfield et al., 2024
Further adapted by Lennart Heim.

"Computing Power and the Governance of AI" Sastry, Heim, Belfield, Anderljung, Brundage, Hazell, O'Keefe, Hadfield, et al. 2024

February 2024

Summing up

- **The leap** from computation (measured in FLOPs) to systemic risk seems arbitrary. I.e. the risks does not necessarily link so strictly to training computation.
- **The flexible** designation of systemic risk is rather discretionary which opens up for other concerns:
  - *What sort* of lawmaking is this? W*hat are the implications* of this lack of clarity and predictability?
  - Drastic interpretation: *A politicised product* devised under intensely stressful negotiations. *A geopolitical* tool?
  - Faithful interpretation: *Necessary iteration with scientific communities and GPAI producers for "future-proof" regulation.*
- **Speculation:** we may need other proxies for systemic risk soon. For example autonomy, personalisation, anthropomorphic design