



Ethics and Regulation of General-Purpose AI Models (GPAI) in the EU



Roadmap

An aerial photograph of a dark, winding road that snakes through a dense, light-colored forest. The road has white lane markings and curves in a series of S-shapes. The forest is composed of many small, coniferous trees, creating a textured, speckled appearance. The overall tone is monochromatic, with the road appearing as a dark line against the lighter forest floor.

1. Introduction: Risks and Governance of GPAI/Foundation Models

1.1 The risks posed by GPAI

1.2 The approach to AI governance: from ethics to regulation

2. The EU AI Act

2.1 General lines of the regulation

2.2 Need and challenges to incorporate GPAI into the Regulation

3. The final approach re GPAI incorporated into the EU AI Act – what we know and what remains open for the future?

(Open discussion and questions)

EU AI Act: first regulation on artificial intelligence

[Society](#) Updated: 14-06-2023 - 14:06

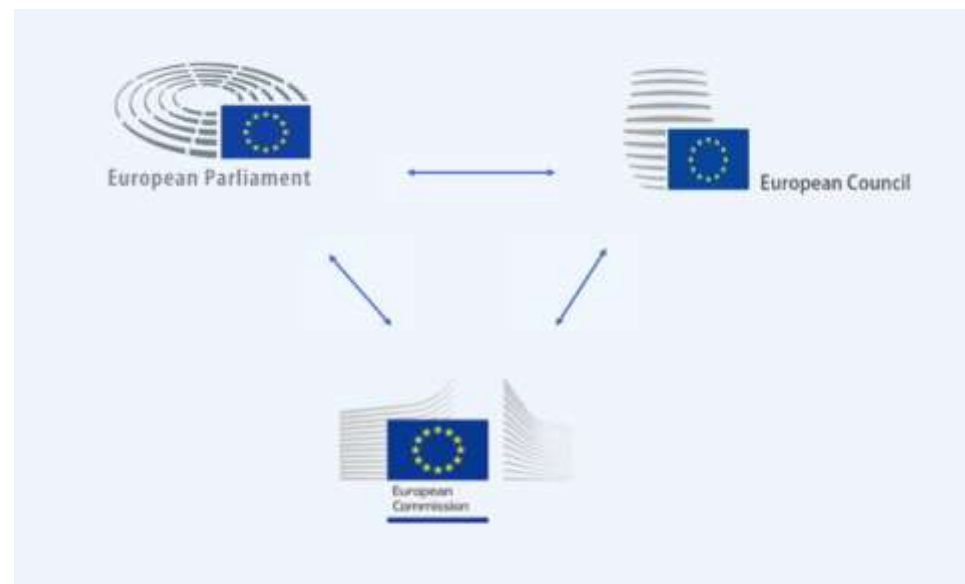
Created: 08-06-2023 - 11:40



The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.



An intense week for AI Regulation in the EU



One of the major roadblocks in the trilogues...

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

By [Luca Bertuzzi](#) | [Euractiv.com](#) ⌚ Est. 6min

📅 Nov 10, 2023 (updated: 📅 Nov 15, 2023)

A technical meeting on the EU's AI regulation broke down on Friday (10 November) after large EU countries asked to retract the proposed approach for foundation models. Unless the deadlock is broken in the coming days, the whole legislation is at risk.


EU: France, Germany and Italy risk unravelling landmark AI Act negotiations



Agreement is reached on Dec 8, after 37h of meetings



Historic!

The EU becomes the very first continent to set clear rules for the use of AI 

The [#AIAct](#) is much more than a rulebook — it's a launchpad for EU startups and researchers to lead the global AI race.

The best is yet to come! 👍

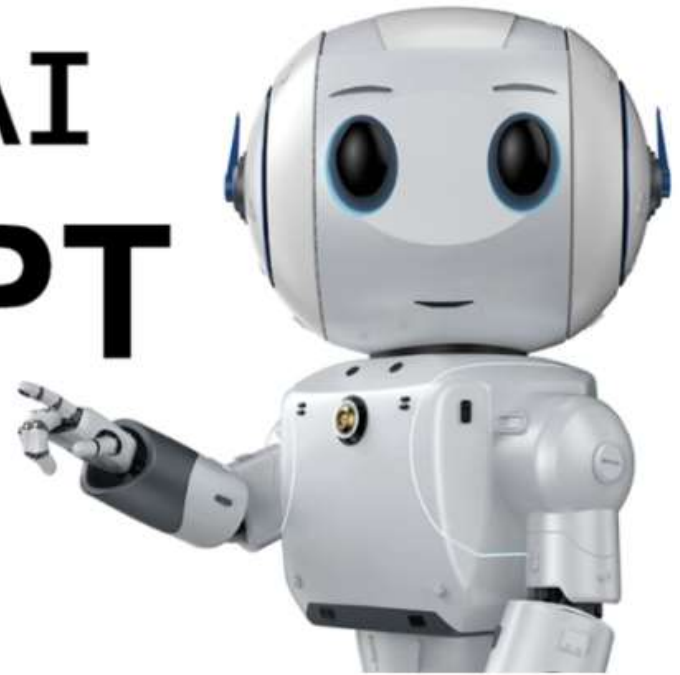


- General lines established
- **Regulation of GPAI models advanced**
- But many details remain to be specified/clarified/corrected in the future regarding regulation of generative AI or, as mentioned by EU, Foundation Models (FM)

**Let us go back
till November 2022, no plans to include rules on generative
AI in the EU AI Act**

**challenges posed by AI
intensified and, for some,
new ones were created**

 **OpenAI**
ChatGPT





Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

TECH

Elon Musk and other tech leaders call for pause on 'dangerous race' to make A.I. as advanced as humans

PUBLISHED WED, MAR 29 2023•8:23 AM EDT | UPDATED WED, MAR 29 2023•11:30 AM EDT



Ryan Browne
@RYAN_BROWNE_

SHARE    



INDEPENDENT

Tech

Artificial intelligence warning over human extinction labelled ‘publicity stunt’

Professor Wachter said the risk raised in letter is “science fiction fantasy” and she compared it to the film *The Terminator*.

She added: “There are risks, there are serious risks, but it’s not the risks that are getting all of the attention at the moment.”

“What we see with this new open letter is a science fiction fantasy that distracts from the issue right here right now. The issues around bias, discrimination and the environmental impact.”

“The whole discourse is being put on something that may or may not happen in a couple of hundred years. You can’t do something meaningful about it as it’s so far in the future.”

“But bias and discrimination I can measure, I can measure the environmental impact. It takes 360,000 gallons of water daily to cool a middle-sized data centre, that’s the price that we have to pay.”

“It’s a publicity stunt. It will attract funding.”

Misinformation, Bias, Discrimination

The Washington Post
Democracy Dies in Darkness

ChatGPT breaks its own rules on political messages

A Washington Post analysis found that the chatbot will draft political messages tailored for demographic groups, like suburban women or rural men



SCI
AM

[Sign Up for Our Daily Newsletter](#)

NOVEMBER 22, 2023 | 3 MIN READ

ChatGPT Replicates Gender Bias in Recommendation Letters

A new study has found that the use of AI tools such as ChatGPT in the workplace entrenches biased language based on gender

Data input

- Use of proprietary databases
- Data protection
- Training of data



THE LEADING LLM MODELS

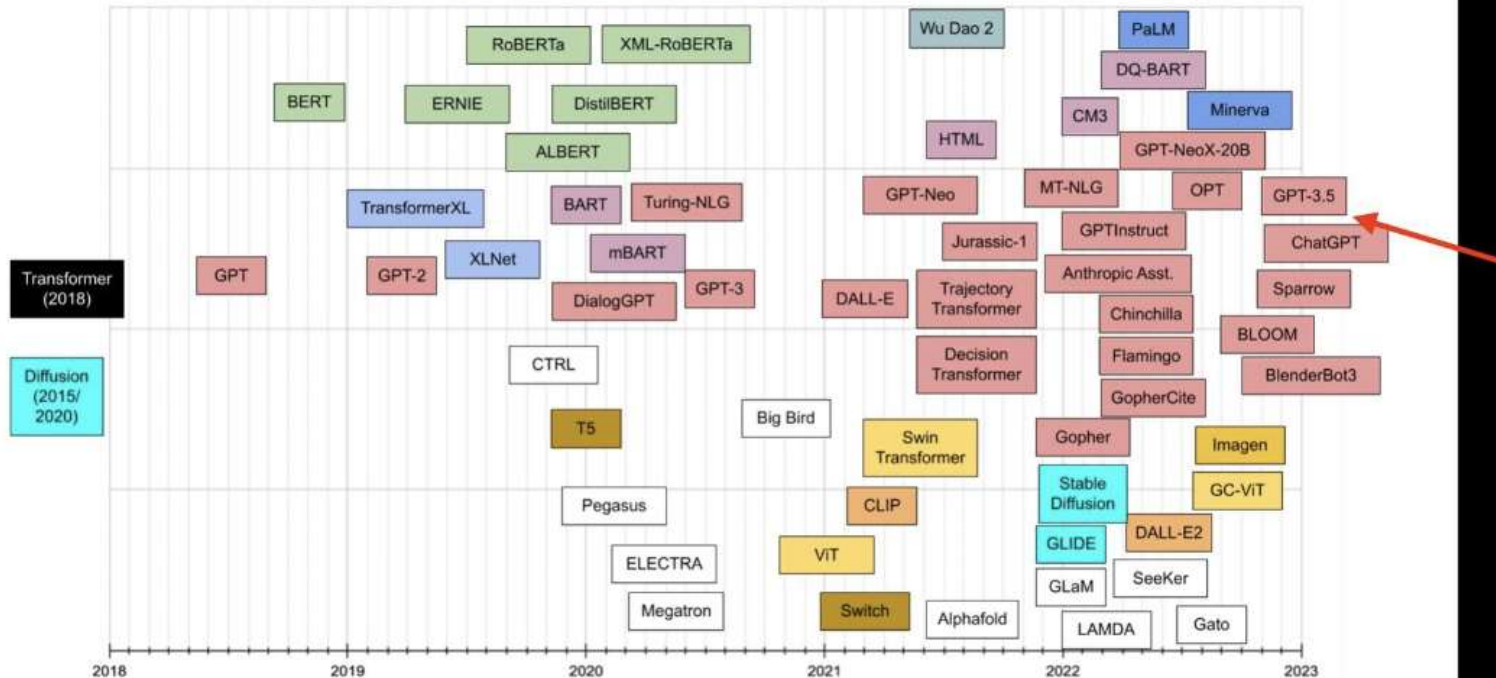
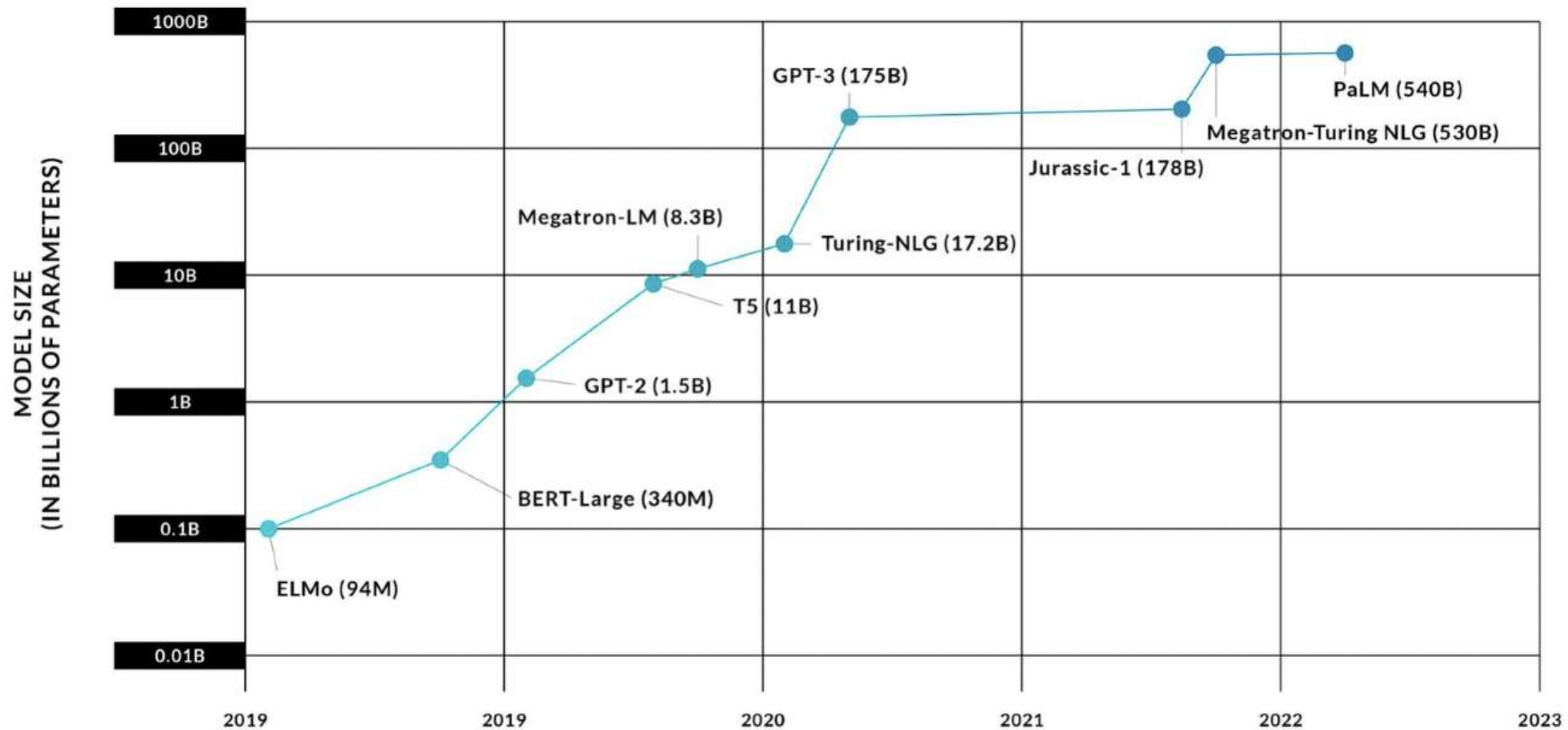


Figure 7: Transformer timeline. Colors describe Transformer family.

XAVIER AMATRIAIN, TRANSFORMER MODELS: AN INTRODUCTION AND CATALOG, ARXIV:2302.07730 (2023) [HTTPS://ARXIV.ORG/ABS/2302.07730](https://arxiv.org/abs/2302.07730)

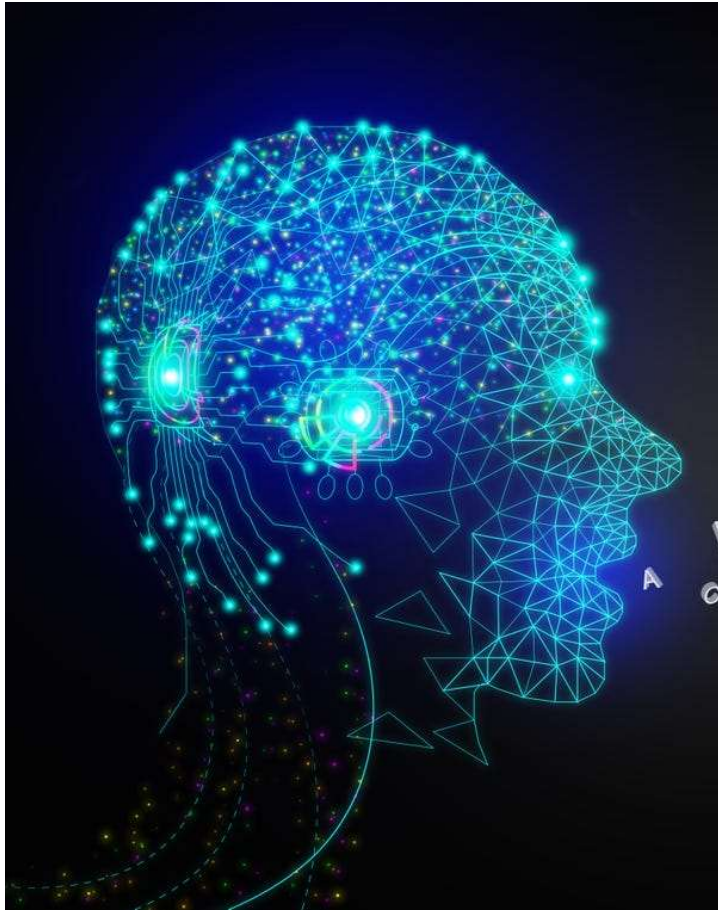
- Daniel Katz, available at <https://speakerdeck.com/danielkatz/generative-ai-plus-law-background-applications-and-use-cases-including-gpt-4-passes-the-bar-exam?slide=205>

Language Model Sizes Over Time



- Daniel Katz, available at <https://speakerdeck.com/danielkatz/generative-ai-plus-law-background-applications-and-use-cases-including-gpt-4-passes-the-bar-exam?slide=205>

Data output



- Who owns, is it protected?
- Liability

“AGI”

Chat about any topic

Answer all your burning questions

Generate realistic images

Do your homework for you

False and misleading information

Propaganda and deception

Biases and hallucinations

Homogeneity and misrepresentation of
language/culture
Harmful and violent content

Gather your data to improve models

Exploitation of underpaid workers

Erosion of rich human practises

Raising the barrier to entry in AI

Tonnes of carbon emissions

Huge quantities of energy/water

Private information

Copyright infringement

Rare metals for manufacturing hardware

<https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/>

Triggered legal/regulatory discussions across the board

- Calls for different regulatory responses in different timelines and magnitudes
 - Short-term (e.g., content moderation issues intensified by bots powered by FM)
 - Medium-term (e.g., algorithmic fairness to ensure equality)
 - Long-time (e.g., existential threat of AI)
- Across different legal fields (product safety, copyright law, anti-discrimination law, data protection law, content moderation...)
- Across different legal instruments (DSA, EU AI Act, copyright laws...)

Global responses

G7 - Hiroshima Process

OECD Policy Consideration

EU

Chinese Law



The first movement: AI Ethics






-
- The rise of regulations may take time (political/legislative process, expertise)
 - Different players want to influence what rules should look like
 - Responsible actors want to protect their consumers & their reputation
 - Try to shield from legal liability

A 3D rendering of a puzzle. The majority of the puzzle pieces are dark grey, while a few on the right side are white. One piece, located in the center-right, is a vibrant red and stands out prominently. The puzzle pieces have a three-dimensional, blocky appearance with visible shadows and highlights.

Need for
enforceable
rules!

The second
movement: AI
Regulation



Rise of Proposals for Regulatory Frameworks on AI

- Different proposals/approaches
 - EU, Brazil, Canada: horizontal, harmonized, human-rights based, enforceable
 - US, UK, Japan: INSO FAR fragmented, sectoral, market-driven, soft compliance
 - China: technical standards and algorithmic management system



EU AI Act



EU AI Act

- **Product safety law**
- **What requirements should AI systems comply with to enter the market?**
- **Post-market monitoring**
- **New legislative framework (NLF)**
 - **General binding rules/principles**
 - **Complemented by technical standards**



Regulation of AI in Europe – Existing Framework

- **Data Protection:** GDPR
- **Automated decision-making:** Art 22 GDPR
- **Safety:** General Product Safety Directive 2001/95; Machinery Directive 2006/42
- **Facial Recognition Systems:** rules restricting export outside EU to countries repressing human rights
- **Human rights:** EU Chart, ECHR
- **Platform Providers:** E-Commerce Directive 2000/31; P2B Regulation 2019/150; DMA 2022/1925; DSA 2022/2065
- **Algorithmic Discrimination:** Anti-Discrimination Directives
- **Algorithmic Manipulation:** Unfair Commercial Practices Directives
- **Liability:** Product Liability Directive 85/577; Digital Content Directive 2019/770; Sale of Goods Directive 2019/771

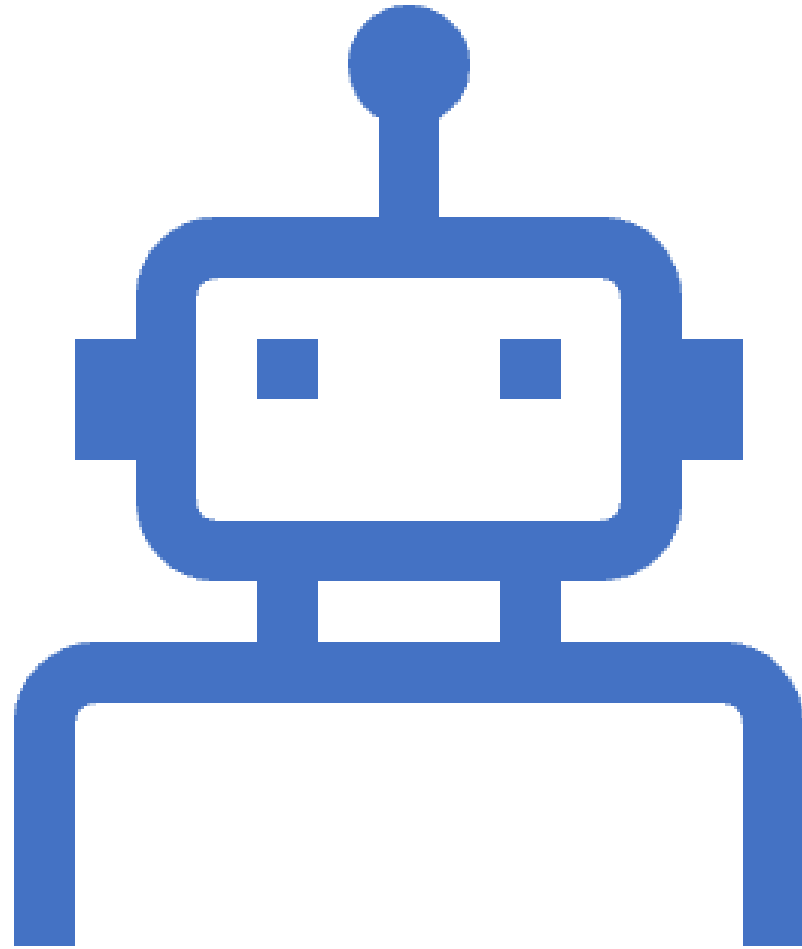


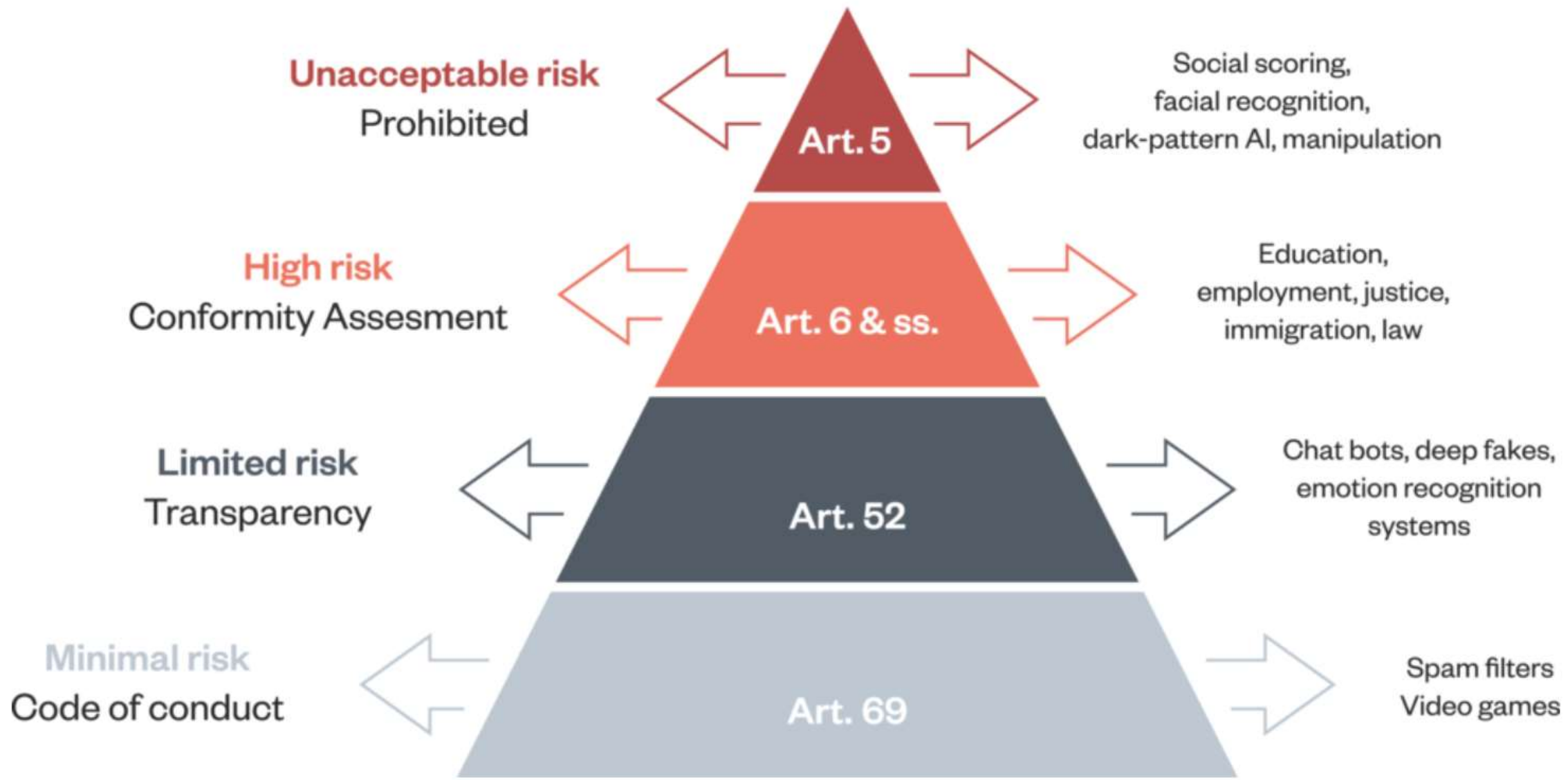
Emerging Regulation of AI in Europe

- Proposal for an Artificial Intelligence Act (2021) 206
- Regulation on Machinery Products, COM (2021) 202
- Regulation on General Products Safety, COM (2021) 346
- New Product Liability Directive, COM (2022) 495
- AI Liability Directive, COM (2022) 496



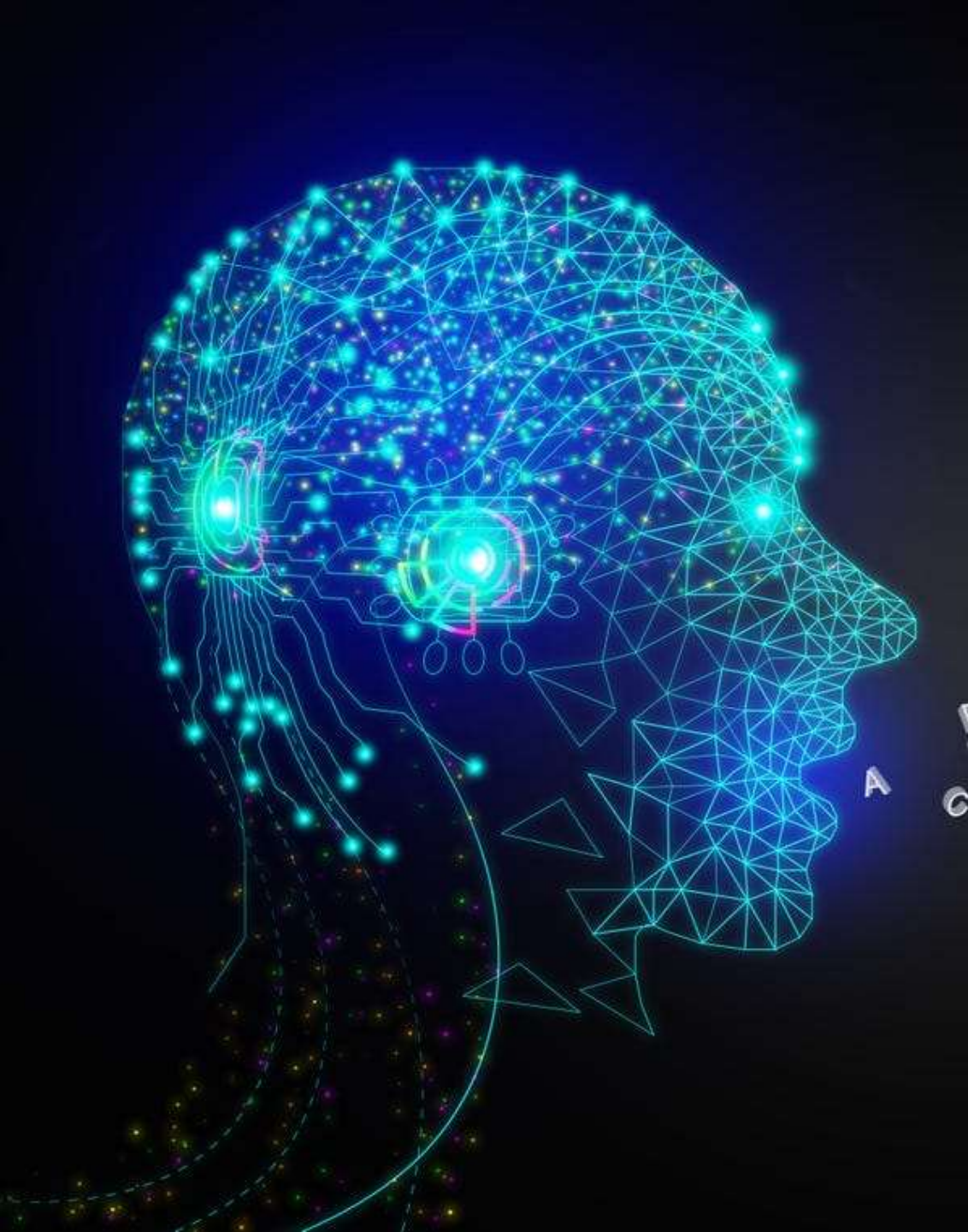
AI Regulation – Risk-Based Approach





General Purpose AI Models

- Foundation models, Generative AI, Large Language Models...
- Generative of Text, Images, Music, Voices, Videos...
- Although they exist for some time, became more intensely subject of regulatory discussions when achieved higher capacity and number of users (initially with ChatGPT3)



On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*¹

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

- Foundation model = “any model that is trained on broad data that can be adapted to a wide range of downstream tasks”
- “From a technological point of view, foundation models are not new — they are based on deep neural networks and self-supervised learning, both of which have existed for decades. However, the sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible”
- “GPT-3 has 175 billion parameters and can be adapted via natural language prompts to do a passable job on a wide range of tasks despite not being trained explicitly to do many of those tasks”
- “have the potential to accentuate harms”

I) Challenges to regulate GPAI under the EU AI Act

1) Should we regulate? Regulation or voluntary codes of conduct?

- Last minute proposal by France, Germany, Italy to avoid regulation of FM and resort instead to voluntary codes of conduct
- Need for European countries to be leaders in technological development, not only the leader in regulation
- SOLUTION: Latest version of the EU AI Act adopted mandatory rules, but with compromises

2) GPAIs are general purpose – but product safety law regulates according to specific purpose

- What are the specific requirements that must be fulfilled then?
- Challenging to perform risk classification is defined according to finality of AI system
- SOLUTION: Tiered approach; different actors have different obligations

3) A Tiered Approach of Regulation

- Tiered approach: distinguish between powerful/systemic risk models from ordinary GPAI
- Not discourage SMEs by excessive burdensome obligations, allowing flourish without excessive regulation
- CHALLENGE: Distinguish Tier 1/2

Established obligations for FM under AI Act

1) Tier 1 – Models presenting systemic risk (Flops 10^{25})

- Floating point operations (FLOPS) is greater than 10^{25}
- “FLOPS refers to the number of operations that a computer can perform in a second, and the 10^{25} number refers to the power of the supercomputer that a model is trained on, and for how long — essentially how much raw computing power has gone into the training process.
- Only ChatGPT4 and Gemini

2) Tier 2 – other foundation models

Tier 1 – FM models with systemic risk

Flops > 10^{25}

- * Risk Management: Organizations must perform model evaluations using state-of-the art protocols and tools.

- * Red Teaming: There is a necessity to conduct and document adversarial testing to identify and mitigate systemic risks.

- * Cybersecurity: Maintaining an adequate level of cybersecurity for both the AI model and its physical infrastructure is non-negotiable.

- * Energy Consumption: Entities must track, document, and report on the known or estimated energy consumption of the model.

Tier 2 – other foundation models

- Subject to transparency obligations
 - Summary of training data
 - Notification when one is interacting with AI system
 - Compliance with EU copyright provisions

Responsibilities of actors in the GPAI Value Chain

