Recent Developments:

The EU AI Act and Policy Design for Large Language Models

Doaa ABU ELYOUNES

Technicalities

The raising hand functionality will be handled through a dedicated website

1) Go to: iraisemyhand.com

2) Enter channel name: **RML2023**

3) Enter your name, and join

Keep the website running in the background and simply press on the raise hand icon any time you have a question/reaction.

The EU AI Act

The AI Act - what seems to be at stake for the EU?

- Optimization, resource allocation, etc. is especially needed in high-impact sectors, including climate change, environment and health, the public sector, finance, mobility, home affairs and agriculture.
- AI brings about new risks
- The EU is committed to strive for a **balanced approach**.
- It is in the Union interest to preserve the EU's **technological leadership**

Twin objectives: Promoting the uptake of AI and of addressing the risks associated with certain uses of such technology. The AI Act seeks to implement the second objective: the development of an ecosystem of trust by proposing a legal framework for trustworthy AI

In the broader EU context

The proposed regulation is part of a tranche of proposals which must be understood in tandem, including:

- The Digital Services Act (with provisions on recommenders and research data access);
- The Digital Markets Act (with provisions on AI-relevant hardware, operating systems and software distribution);
- Announced product liability revision relating to AI
- The draft Data Governance Act (concerning data sharing frameworks)

The AI Act - specific objectives

- Ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values
- Ensure legal certainty to facilitate investment and innovation in AI
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems
- Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation

Key elements

- A single future-proof definition of AI
- Proportionality - imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety. (except transparency)
- Risk-based approach / methodology to define 'high-risk" AI
- Rertain particularly harmful AI practices are prohibited as contravening Union values,
- Specific restrictions and safeguards are proposed in relation to certain uses of remote biometric identification systems for the purpose of law enforcement.
- Throughout the whole AI systems' lifecycle
- Transparency obligations (flag AI is being used)

Other elements

- Reporting obligation for high-risk AI applications in a public EU-wide database + inform incidents

The Definition

Article 3 Definitions

For the purpose of this Regulation, the following definitions apply: (1) ‘artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;

ANNEX I - ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES referred to in Article 3, point 1

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

Art. 5 Four levels of risk

i) Unacceptable risks (Title II)

ii) High risks (Title III)

iii) Limited risks (Title IV)

iv) Minimal risks (Title IX).

Title II: Prohibited practices (Unacceptable risks)

- Manipulative systems
  - subliminal systems
  - “in order to” materially distort behavior
  - and **individual** physical or psychological
- Social scoring
  - by or on behalf of public authorities
  - leading to…
- Biometric systems in publicly accessible spaces **by law enforcement**
  - 3 exceptions
  - Excludes public space *online*! (*see* recital 9)
  - No “placing on the market” prohibition

Title III: High risk systems (Art. 6 - annex II and annex III)

- AI systems that are products or safety components (broadly construed) of products already covered by certain Union health and safety harmonisation legislation (such as toys, machinery, lifts, or medical devices).
- ‘Standalone’ AI systems specified in an annex for use in eight fixed areas: (Comes from product regulation)

- Biometric identification - remote and ‘post’ (v. art. 5)
- Management and operation of critical infrastructure
- Educational and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential services and benefits
- Law enforcement
- Migration, asylum and border management
- Administration of justice and democracy

Title IV: Limited risk

- Concerns systems that are (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content (‘deep fakes’)
- Mainly transparency requirements
- When persons interact with an AI system or their emotions or characteristics are recognized through automated means, people must be informed of that circumstance
- If an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means

Title IX: Minimal risk

- Relevant to all other AI systems
- Creates a framework for the creation of codes of conduct, which aim to encourage providers of non-high-risk AI systems to apply voluntarily the mandatory requirements for high-risk AI systems (as laid out in Title III)
- Self implementation
- Those codes may also include voluntary commitments related, for example, to environmental sustainability, accessibility for persons with disability, stakeholders' participation in the design and development of AI systems, and diversity of development teams

What do you think about the risk based approach?

Governance and implementation

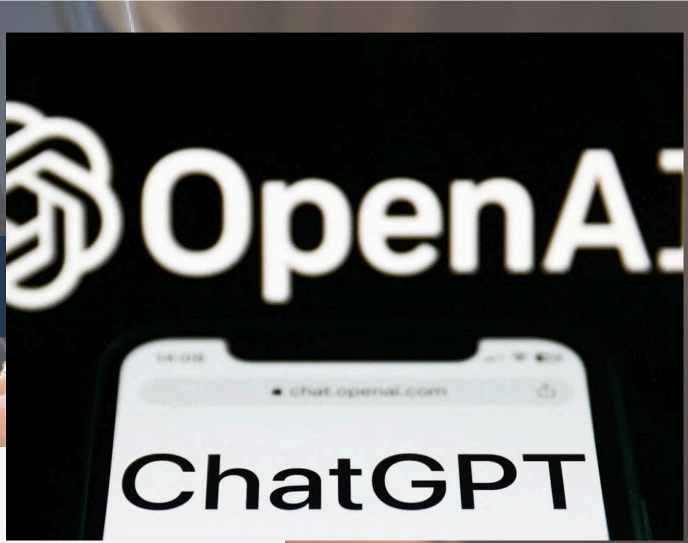
- Member States are required to designate coordinating agencies for the implementation of the act
- An AI Office within the Commission is set up tasked to oversee the most advanced AI models, contribute to fostering standards and testing practices, and enforce the common rules in all member states
- The AI Board, which would comprise member states’ representatives, will remain as a coordination platform and an advisory body to the Commission and will give an important role to Member States on the implementation of the regulation
- An advisory forum for stakeholders, such as industry representatives, SMEs, start-ups, civil society, and academia, will be set up to provide technical expertise to the AI Board

Policy Design for Large Language Models

Introduction

- What are the technical characteristics that make LLMs and other foundational models unique?
- Are these unique characteristics merit special treatment in terms of policy design?
- Do you think it will be easy/ possible to apply the policy considerations we studied with regard to fairness, privacy and transparency to LLMs? If changes are required what kind of changes?

RECORD SPEED OF UPTAKE: ChatGPT



Time taken to reach 1m users (mths)

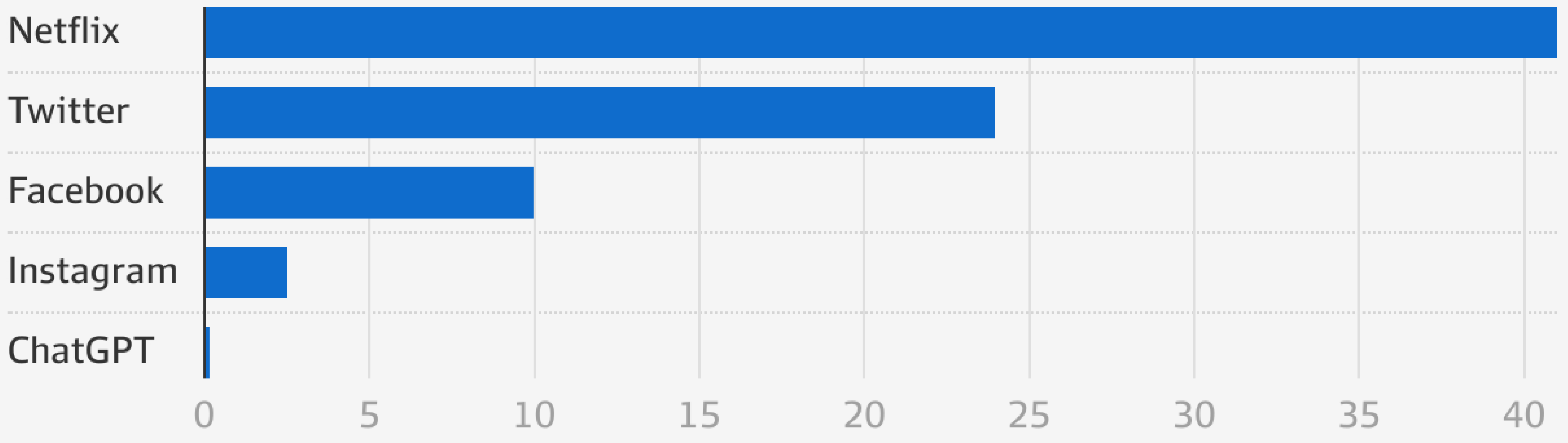


Chart: Financial Review • Source: Genevieve Roch-Decter, CFA

Benefits of foundational models

- Potential to increase productivity
- Improve access to a wide range of services for people in economically and socially disadvantaged positions
- Enhance and personalize user experiences, particularly in education and healthcare
- Improve decision making

Fairness and nondiscrimination

- Stereotypical and discriminatory outputs
  - Racial discrimination- describing white and Asian men as better scientists; or classifying US and Canadian workers as "senior“, and Mexican workers as "junior
  - Gender biases- depicting female characters as less powerful and defining them by their physical appearance and family roles
  - Religious biases- researchers from Stanford found that Muslims were depicted as terrorists in 23% of the prompt they tested, while Jews were associated with money in 5%
- It is more challenging to foresee and check for biases
- Lack of linguistic diversity

Transparency, explainability and verifiability

- The models are often opaque both in relation to the data set that has been used to train them and the workings of the system itself in how it derives its answers
- The outputs produced are often not accurate or up to date, furthermore, when prompted to provide references or citations, they often fabricate made-up resources to support their outputs
- Lack of transparency and verifiability could contribute to the spread of disinformation and misinformation as well as to the creation of deep fakes

Accountability and human in the loop

- The terms of use of foundational models delegate responsibility entirely to the users
- Copyrighted materials and whom responsibility should be delegated to in case of an error
- Stability AI (the company that makes the AI tool Stable Diffusion) is currently being sued by Getty Images for copyright infringement
- There are many other lawsuits in the pipelines claiming copyright infringement
- Some argue that humans could be deemed as creators of inventions produced by AI with adequate human supervision, however, the definition of "adequate" remains unclear
- The fair use doctrine is not guaranteed for foundation models as they can generate content that is not "transformative" enough compared to the copyrighted material

Privacy and data protection

- With the right prompt, foundational models could reveal data from their training data set, including providing personal information about individuals collected from the open internet that may never have been intended to be processed and made available in this way and within this use context.
- For example, in March 2023, ChatGPT was briefly taken offline after experiencing a bug that allowed some users to see the titles from another user's chat history and may also have made visible the payment information of some subscribers
- In March 2023 the Italian data protection authority blocked the use of ChatGPT, citing privacy concerns about the way it was gathering data as well as its lack of age verification, and opened an investigation into whether the tool is compliant with the General Data Protection Regulation. This block is lifted but investigation is on going

Safety and security

- Foundational models can be used to facilitate the distribution of content that may, by its very nature, be dangerous. For example, through a particular prompt, ChatGPT has been able to deliver instructions for building a dirty bomb
- The tools' programming capabilities can be used to facilitate the creation of computer viruses, including malware, ransomware, spyware and phishing campaigns
- The very thing that makes foundational models so good—the fact they can follow instructions—also makes them vulnerable to being misused. That can happen through "prompt injections," in which someone uses prompts that direct the language model to ignore its previous directions and safety guardrails

Other concerns

- Impact on the environment: a comprehensive evaluation of the environmental impact of generative AI is yet to be provided
- The degradation of social interactions: mediating human communication is particularly relevant in education, where the teacher student relationship is at risk
- Harming critical thinking and creativity: writing could cease to be a creative and reflective space if the practice of editing text created by AI were normalized without safeguards in place
- Impact on the economy and labor: foundation models increase concerns about the impact of AI on labor markets, and the speed and depth with which certain jobs will be transformed. These tools can be used to automate tasks traditionally associated with human functions that include reasoning, writing, creating graphics, and analyzing data

Regulating foundational models through the EU AI Act

- The EU converged on minimum standards for all foundation models (called general-purpose AI models)
- Providers of general purpose AI models will be obliged to maintain technical documentation and provide sufficient information about their model so that downstream providers that incorporate it into their system can comply with their own AI Act obligations
- General purpose AI model providers will also be required to have a policy concerning the respect of EU copyright rules, particularly to ensure that, where copyright holders have opted out of allowing their data to be available for text and data mining (including web-scraping), this is identified and respected.
- Providers must also prepare and publish a statement about the data used to train the general purpose AI model
- The Act defined also more stringent rules for so-called high-impact foundation models with systemic risk
- Such strict rules kick in if the model was trained with more than 10^25 FLOPs (floating-point operations, roughly equivalent to calculation steps)

Conclusion

- Foundational models are creating high expectations of the services they can provide to humanity, however, their widespread use is also highlighting the risks attached to how these technologies are currently being deployed, responding to a frantic technological race between economic actors and countries, instead of serving the public good. To get it right, we need the right oversight and policy frameworks
- Foundational models are often described as “experimental” by their developers, and it is often only after they have been released to the public that harms start to become apparent, even when these could and should have been anticipated at the design and development stages
- Ethical considerations and processes to support them must be built into every stage of the life cycle of such models, in an ex-ante manner to identify and address risks effectively, and to prevent ethics being sidelined while other considerations such as commercial or economic competition prevail

Thank you