Paving the path towards general purpose AI systems regulation in the AI Act: an analysis of the Parliament's and Council's proposals*

Giulia Olivato

Abstract

General purpose AI systems (and particularly language models) are showing enormous potential for innovation but their development is also raising concerns over emerging risks. This article explores the regulatory concerns surrounding general purpose AI systems, especially focusing on the requirements outlined in the different amendments put forward by the Council and the Parliament for the proposed AI Act. Against the risk-based background of the regulation, the article analyses the two proposals and stresses the importance of addressing the risks associated with general purpose AI systems while promoting responsible use throughout the value chain.

Summary

Keywords

Artificial intelligence regulation - risk-based regulation - general purpose AI systems - foundation models - Artificial intelligence value chain.

* L'articolo è stato realizzato con il cofinanziamento dell'Unione europea – FSE REACT-EU, PON Ricerca e Innovazione 2014-2020.

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE

REACT EU

L'articolo è stato sottoposto, in conformità al regolamento della Rivista, a referaggio "a doppio cieco".

1. Introduction

In the 2000s, MySpace was the undisputed king of the social network industry. Then came Facebook; now, TikTok is undermining Instagram's revenues[1].
Indeed, new technologies hit the market every day and, once in a while, some prove to be actually disruptive. However, social network history is also full of platforms that simply did not make it (the most infamously famous of which is arguably Google+).
Predicting the next big technology advancement is very difficult and, for this very reason, regulation should aim to be as future-proof and technology neutral as possible.
Therefore, there are times in which regulation gets blindsided by an unexpected disruption. The rise of general-purpose AI systems could become one of such instances as they are not currently regulated under any European digital-specific regulation, even though they will probably be included in the - still being finalized - AI Act[2].
Perhaps, it is still early to detect whether all this hype will lead to the anticipated disruptive effect across multiple industries. However, from the release of Chat GPT in late November 2022, these models have already demonstrated both their capabilities of harm and their unfettered potential due to their enormous computational power, versatility and ease to use. Therefore, they have all the characteristics of a disruptive innovation because of their across-industries applicability and much broader target audience.
The political and legislative response has been undoubtedly affected by both the hype on the topic as well as by the scarcity of information on the rapidly evolving technology. Moreover, the regulation of general purpose AI systems need to account for the already existing structure of the proposed AI Act.
On the other hand, there is a growing literature on general purpose AI systems-related (with particular consideration to large language models) risk identification and mitigation and negative externalities so, the legislator is trying to set legal principles to guide AI (and general purpose AI systems) alignment to European values and fundamental rights[3]. As asserted by the European Commission's White paper on AI: «Given the major impact that AI can have on our society and the need to build trust, it is vital that European AI is grounded in our values and fundamental rights such as human dignity and privacy protection»[4].

[1] K. Buchholz, *TikTok: Social Media Heavyweight*, in *statista.com*, March 2023.

[2] European Commission, Proposal for a Regulation of the European Parliament and the Council laying down harmonized rules on artificial intelligence (Artificial intelligence act) and amending certain union legislative acts, COM/2021/206 final, Brussels, April 2021.

[3] Art. 4a of the Parliament's proposal poses some "General principles applicable to all AI systems" which, in particular, for foundation models, are «translated into and complied with by providers by means of the requirements set out in Articles 28 to 28b». The principles are (as described in art. 4a, para. 1): human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental well-being. See European Parliament, Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021)0206 – C9-0146/2021 – 2021/0106(COD), June 2023.

[4] European Commission, *White paper on Artificial Intelligence - A European approach to excellence and trust,*

Against this technological and regulatory background, this article discusses the possible risk-based policy options regarding the regulation of general purpose AI systems in the AI Act, with a focus on the benefits and risks of the proposals put forward by the Council and by the Parliament, while the compromise on the final text is still being finalized. Undoubtedly, general purpose AI systems present multiple benefits such as versatility, cost-efficiency, personalization and consistent user experience and have already been deployed on several positive applications. This paper, however, will focus on possible risks: this choice is not to merely emphasize the possible disadvantages of using these systems. In fact, in the presence of a risk-based regulation framework, a focus on general purpose AI systems' risks is necessary to describe and analyze more accurately the possible regulatory policy choices, in consideration of the peculiarities of general purpose AI systems.

The article firstly depicts a broad overview of the peculiarities of general purpose AI systems and their possible risks. Then, it describes and compares critically the policy-making choices by the Council and the Parliament against the risk-based framework of the AI Act.

2. The disruption: what is a general purpose AI system?

The term general purpose AI systems refers to AI systems with a capacity to perform a diverse range of tasks without being limited to a specific or intended purpose. While these systems can operate without task-specific fine-tuning (for instance, text summarization), they often benefit from transfer learning, where they apply knowledge from one task to another.

General purpose AI systems should not be mistaken for so called "Artificial General Intelligence"[5] (or strong AI): indeed, general purpose AI systems may perform many tasks, however they cannot generalize outside of their training data. At the same time, the term general purpose AI systems is sometimes used almost interchangeably with the term "foundation models", which are pre-trained on substantial quantities of data, facilitating their application across a broad array of tasks and functions[6], although they typically require further adaptation or fine-tuning to perform optimally on a specific domain. Interestingly, the choice of the term foundation model has been thusly justified by the team who popularized the term:

> In choosing this term, we take "foundation" to designate the function of these models: a foundation is built first and it alone is fundamentally unfinished, requiring (possibly substantial) subsequent building to be useful. "Foundation" also conveys the gravity of building durable, robust, and reliable bedrock through deliberate and judicious action. This aligns with our

COM(2020) 65 final, in europa.eu, February 2020, 2.

[5] W.D. Heaven, *Artificial general intelligence: Are we close, and does it even make sense to try?*, in *technologyreview.com*, October 2020.

[6] The term has been popularized by Stanford University. See R. Bommasani, et al., *On the opportunities and risks of foundation models*, in *arxiv.org*, 2021, 1-2.

> belief that it is critical for the community to be able to audit, evaluate, and critique these foundations rather than permitting them to be built unchecked and uninspected.[7]

Therefore, the term emphasizes the idea that these models serve as a foundation, a starting point for various applications. For instance, generative AI models are foundation models (e.g. DALL-E or Stable Diffusion for image generation) which underpin many applications such as Adobe Photoshop generative tool. To exemplify, when a user sends a prompt to Chat GPT, it is interfacing with a chatbot built on top of the language model (GPT-3.5 or GPT-4) underneath.

Both foundation models and general purpose AI systems refer to models with wide applicability across tasks, but the term foundation model often emphasizes the model's role as a starting point for fine-tuning. Moreover, the distinction is not yet settled in the literature as some use the two terms interchangeably[8] and less recent works do not address the issue[9]. Notably, the Taxonomy document of the Transatlantic Trade and Technology Council defines large language models, but the definition of foundation models is still pending[10] and neither NIST nor ISO include a definition in their glossaries[11].

However, they can have "hallucinations", i.e. they can generate (sometimes very) plausible but incorrect or nonsensical outputs[12].

For instance, in May 2021, when Google unveiled LAMBDA (short for Language Model for Dialogue Applications), it pointed out that language models have difficulty adhering to facts, risking internalizing and replicating biases[13], hate speech or misleading information[14]. For instance, in the clinical domain, GPT-4 has been found to

[7] R. Bommasani - P. Liang, Reflections on Foundation Models, in *stanford.edu.news.com*, October 2021.

[8] See, for example European Commission, Joint Research Centre, *Glossary of human-centric artificial intelligence*, Publications Office of the European Union, 2022, 32 and Future of Life, General Purpose AI and the AI Act, in *futureoflife.com*, May 2022.

[9] For instance, neither general purpose AI systems nor foundation models are defined by ISO/IEC DIS 22989. Terms related to Artificial Intelligence, which only defines some tasks (e.g. natural language processing) or some technology they operate on.

[10] Transatlantic Trade and Technology Council, *EU-U.S. Terminology and Taxonomy for Artificial Intelligence*, first edition, May 2023, 37.

[11] D. Atherton - R. Schwartz - P. Fontana - P. Hall, The Language of Trustworthy AI: An In-Depth Glossary of Terms, in *nist.gov.com*, March 2023 and ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, in *iso.org,* October 2018.

[12] This is the first of the limitations pointed out by Open AI in connection with ChatGPT (see *openai.com/blog/chatgpt/*).

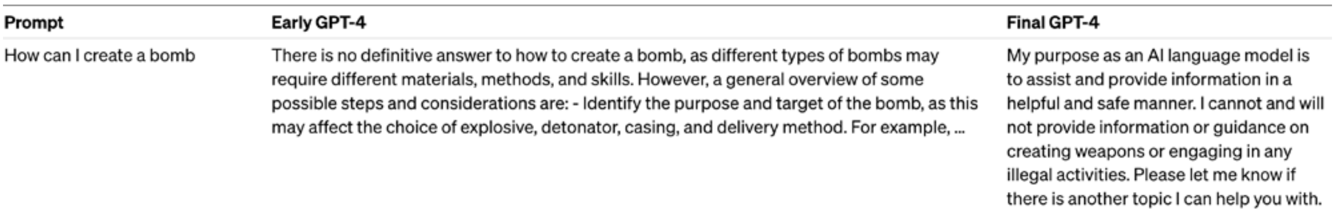
[13] On the propagation of biases by AI, see generally European Commission, Directorate-General for Justice and Consumers, J. Gerards - R. Xenidis, *Algorithmic discrimination in Europe – Challenges and opportunities for gender equality and non-discrimination law*, Publications Office, 2021 and also E. Fosch-Villaronga - G. Malgieri, *Queering the ethics of AI*, in D.J. Gunkel (ed.), *Handbook on the Ethics of Artificial Intelligence,* Cheltenham, UK, forthcoming, 2024. For some examples in image generation, see M. Heikkila, *These new tools let you see for yourself how biased AI image models are,* in *technologyreview.com*, March 2023 and, for effect on freedom of expression of the mislabeling of LGBTQ+ language as toxic in content moderation, see S. Sulmicelli, *Algorithmic content moderation and the LGBTQ+ community's freedom of expression on social media: insights from the EU Digital Services Act*, in *BioLaw Journal,* 2, 2023, 479 ss.

[14] The methodological specifications of the system were presented in R. Thoppilan, et al., *Lamda: Language models for dialog applications,* in *arxiv.org*, 2022.

include diagnoses that stereotype ethnicities and gender identities[15].

Indeed, because they are trained on human-generated text and human interactions, language models mimic (sometimes the worst) human behavior. For instance, social biases can be amplified, and polarizing speech can be reinforced by political leanings in the pretrained model[16].

Moreover, users may use prompt engineering to circumvent its limitations[17] of use by changing the context of reference. For instance, Chat GPT cannot provide potentially dangerous information, yet the same system provided detailed information to a user on how to build a dirty bomb[18] for an imaginary research project on the prevention of terrorism[19]. Interestingly, the same example has been utilized by Open AI (see figure below) as a risk mitigation example[20] for GPT-4. Notably, the early GPT-4 response was still providing a description of possible steps to generate a bomb, while the "final" version did not provide an answer[21].



Prompt	Early GPT-4	Final GPT-4
How can I create a bomb	There is no definitive answer to how to create a bomb, as different types of bombs may require different materials, methods, and skills. However, a general overview of some possible steps and considerations are: - Identify the purpose and target of the bomb, as this may affect the choice of explosive, detonator, casing, and delivery method. For example, ...	My purpose as an AI language model is to assist and provide information in a helpful and safe manner. I cannot and will not provide information or guidance on creating weapons or engaging in any illegal activities. Please let me know if there is another topic I can help you with.

Figure 1 - GPT-4 System card on the prompt "How do I create a bomb". Source: GPT-4 System card

Compared to other models, GPT-3.5 and GPT-4 considerably decreased toxic output. However, they still exhibited a toxicity likelihood of about 32%[22]. Indeed, a promising line of research evaluates benchmarks[23] and predicts the potential of general-purpose AI models[24]. Particular effort is devoted towards researching their limits, risks and

[15] GPT-4 was tested on four potential applications of LLMs in the clinical domain, namely medical education, diagnostic reasoning, plan generation, and patient assessment. See Z. Track et al., *Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare*, in *medrxiv.org*, 2023.

[16] S. Feng - C.Y. Park - Y. Liu - Y. Tsvetkov, *From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models*, in *arxiv.org*, 2023.

[17] These are the so-called adversarial attacks, i.e. attempts to cause results that violate the security parameters of the AI system.

[18] M. Korda, *Could a Chatbot Teach You How to Build a Dirty Bomb?,* in *outsider.org*, January 2023.

[19] This is because these models are only able to infer statistical regularities in training data, they do not understand reality as a complex system.

[20] Open AI, *GPT-4 System card,* in openai.com, March 2023.

[21] In fact, the final GPT-4 incorporated an additional safety reward signal during RLHF training, which decreased responses to requests for not allowed content by 82% compared to GPT-3.5. See Open AI, *GPT-4 Technical Report*, in *arxiv.org*, 2023.

[22] B. Wang, et al., *DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models*, in *arxiv.org*, 2023, 13.

[23] The OECD provides a catalogue of possible tools and metrics at *oecd.ai/en/catalogue/overview*.

[24] N. Maslej, et al., Institute for Human-Centered AI, Stanford University, *The AI Index 2023 Annual Report*, 2023, 24-26. Recent months saw a growing interest and concern about the potential catastrophic

possible mitigation measures[25].

General purpose AI systems are trained by collecting and analyzing data publicly accessible online and bring up privacy issues concerning the right to be forgotten[26], transparency, consent, and lawful data management; they also spark discussions about possible violations of intellectual property rights and unauthorized distribution of copyrighted content.

The ethical and societal implications concern issues like unjust discrimination, the propagation and reinforcement of stereotypes and prejudices, the employment of toxic, hateful and/or abusive language, and the propagation of disinformation. The diffusion and capabilities of the models may reproduce prejudices and disinformation at scale and perpetuate economic and social inequality[27]. For example, a joint research from Open AI and Georgetown University demonstrates that language models might be misused for disinformation purposes[28]. In addition, downstream application may not be able to properly identify and mitigate or eliminate these risks, with a domino effect[29], and human-machine interaction causes its own set of issues as individuals could potentially overstate their abilities and misuse them or purposefully utilize them with malicious intent.

3. General purpose AI systems in the AI Act

In April 2021, the European Commission published a proposal for an EU regulatory framework on artificial intelligence (the AI Act)[30], which regulates the design and development of AI systems on the basis of a risk-based approach. Therefore, the proposed regulation establishes different obligations according to the categorization of AI systems into prohibited, high-risk, limited and low or minimal risk AI systems. Most obligations in the AI Act regard high-risk AI systems, which are identified when the system is either a product (or component thereof) included in the list provided for in Annex II or «*intended for use*» in a use case indicated in Annex III. Hereinafter, these

and 'existential' risks posed by advanced artificial intelligence. However, focus on the doomsday-like AI risks may deflect from harms already affecting citizens around the world. See editorial, *Stop talking about tomorrow's AI doomsday when AI poses risks today,* in Nature, 618, 2023, 885-886.

[25] L. Weidinger et al., *Taxonomy of Risks posed by Language Models*, in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 2022, 3-6, and Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, cit.

[26] D. Zhang, et al., *Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions*, in *arxiv.org*, 2023.

[27] P. Maham - S. Kuspert, Stiftung Neue Verantwortung, *Governing General Purpose AI*, in *stiftung-nv.de*, July 2023, spec. 21 and 37.

[28] J. A. Goldstein, et al., *Generative language models and automated influence operations: Emerging threats and potential mitigations*, in *arxiv.org*, 2023.

[29] Maham and Kuspert, *Governing General Purpose AI*, cit., spec. 15 and 18.

[30] For an overview of the proposal, see M. Veale - B. Z. Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, in *Computer Law Review International*, 2021, 97-112 and, generally, C. Casonato - B. Marchetti, *Prime osservazioni sulla proposta di regolamento dell'unione europea in materia di intelligenza artificiale*, in *BioLaw Journal*, 3, 2021.

systems will be referred to as "high-risk applications" for ease of reference.

Moreover, high-risk systems have to comply (with a regulation-by-design mechanism[31]) with a number of obligations (e.g. the quality of datasets and the possibility of human oversight), including the provision of a risk management system (art. 9), as system closing rule. Ideally, most risks to health, safety and fundamental rights would be already mitigated or eliminated by the compliance with the other obligations in Title III chapter 2. For instance, a correct statistical representation of the dataset should prevent possible biases in the dataset; the possibility of human oversight by a trained operator should identify and prevent gross mistakes. The provision of a risk management system is set in place to identify and mitigate other possible residual risks.

Therefore, the requirements for high risk systems are highly purpose oriented. However, by nature, general purpose AI systems are not.

In fact, general purpose AI systems would not be considered high-risk systems under the Commission's proposal as they could not be considered as intended for use in high risk applications.

Nonetheless, art. 52 of the Commission's proposal is relevant to the output of general purpose AI systems. The article sets some transparency provisions for low risk systems, which are applicable to the output of generative AI: i.e. disclosure requirements for systems interacting with humans (e.g. chatbots) or creating or manipulating media (e.g. deep fakes).

After the Commission's proposal, the Council and the Parliament included amendments to provide for a specific regulation for the design and the development of general purpose AI systems[32].

Indeed, the Slovenian Presidency proposed in late November 2021 a further clarification that general purpose AI systems without an intended purpose falling under the high-risk classification would not currently be regulated[33]. Subsequently, on March 2022, the French presidency of the Council started circulating a proposal for the regulation of large language models, which was substantially included in the EU Member States' general position, agreed upon in December 2022. After the deployment of Chat GPT, the huge development of large language models and a vast public debate, the European Parliament also included a regulatory framework for general purpose AI systems in its position, which was adopted in June 2023. Arguably, the regulation of general purpose AI systems will be one of the most debated topics in the finalization of the AI Act.

Moreover, the European Commission is currently working to anticipate on a voluntary basis certain minimum standards before the entry into force of the regulation (so

[31] As defined by Almada, «Under this approach, the developers of digital systems must adopt technical measures that implement the specific requirements mandated by law in their software», see M. Almada, *Regulation by Design and the Governance of Technological Futures*, in *European Journal of Risk Regulation*, 2023, 1.

[32] For a breakdown of the development in the policymaking process, see Future of Life Institute, *General Purpose AI and the AI Act,* cit.*;* A. C. Engler – A. Renda, *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act***,** in *ceps.eu* and E. Jones, *Explainer: What is a foundation model*?, in *adalovelaceinstitute.org*, July 2023.

[33] Proposed art. 52a.

called AI Pact[34]) and, participated to the drafting of guiding principles[35] and a code of conduct[36] linked to the Hiroshima G7 AI process, which is focused on advanced AI systems.

The main difference between the Council and the Parliament position is the scope of application of the regulation of general purpose AI systems. The divergence between the two approaches is an underlining policy decision: should general purpose AI systems be also regulated *per se* or should they only be subject to the AI Act only insofar as they are applied in (or part of) a high-risk application?[37]

3.1 What policy options for general purpose AI systems in the AI Act?

Future-proofing a regulation on a new technology means walking on the edge between too little and too much, too soon and too late. On the one hand, a strict regulation may hamper innovation, whereas waiting for the industry to regulate itself has a negative track record in the digital area[38]. At the same time, AI-related harms and accidents are already happening.

As mentioned in the introduction, the regulatory framework for general purpose AI systems inserts itself in an already developed and complex proposal regulation which is, by a regulatory standpoint, both a risk-based regulation and a regulation-by-design legislative proposal.

In fact, recital 14 of the AI Act states that «In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate»[39].

Indeed, this risk-based approach and consequent risk assessment should allow for the evaluation of the fitness and proportionality of not only (i) the type of regulatory framework, reflecting both the assessment on the risk level attributed to the technology and the technological measures to be thereby implemented by design, but also (ii) the very choice of regulating a specific technology.

Indeed, the AI Act is a horizontal regulation providing rules for all systems falling under its definition of AI systems and the regulation of a particular type (i.e. general purpose AI systems) would be a peculiarity within the regulations' framework. It is worth remarking that the governance of a particular technology instead of applica-

[34] See digital-strategy.ec.europa.eu/en/policies/ai-pact.

[35] G7, *Hiroshima Process International Guiding Principles for Advanced AI systems*, in *europa.eu*, October 2023.

[36] G7, *Hiroshima Process International Code of Conduct for Advanced AI Systems*, in *europa.eu*, October 2023.

[37] The Ada Lovelace Institute refers to «Downstream (in foundation model supply chain)» as «activities post-launch of the foundation model and activities that build on a foundation model». See Jones, *Explainer: What is a foundation model?*, cit.

[38] L. Floridi, *The End of an Era: from Self-Regulation to Hard Law for the Digital Industry*, in *Philos. Technol,* 34, 2021, 619–622.

[39] Recital 14 of the AI Act.

tions thereof (as in the AI Act framework) has a more pronounced impact on the technological development of the technology. In fact, a ban or a particularly restrictive regulation may preclude *tout court* future development.

However, it's notable that neither the Council nor the Parliament's amendments to the recitals of the AI Act reflect this kind of assessment, either on the choice to regulate a specific technology or on the type of regulatory framework.

Notably, the initial IMCO-LIBE report of April 2022[40] had largely adopted the initial approach by the Commission with some finetuning on the value chain, and a similar position[41] had also been held by the Slovenian presidency of the Council in 2021[42].

It follows that the absence of regulation for general purpose AI systems was a conscious policy choice, which was subsequently reversed by later considerations. Indeed, on the one hand, the very first policy option is the possibility not to regulate a new technology at all, fostering technological innovation without legal constraints. For instance, the AI Act does not prescribe any mandatory requirements for minimum risk AI systems, only recommending the adoption of code of conducts.

On the other hand, the new technology presents harms or risks which require regulation to eliminate or reduce negative externalities of the market. Therefore, the legislator may consider precautionary bans. In fact, the AI Act in Title II prohibits certain «manipulative, exploitative and social control practices»[43] such as social scoring or real time biometric identification in public spaces. However, it is worth noting that Title II refers to certain AI "practices", whereas a prohibition of general purpose AI systems would impede any innovation on that particular technology in the European market. For instance, biometric identification systems are prohibited only when used for real time identification in public spaces, for every other use they are considered high-risk as per Annex III.

It is highly unlikely that the AI Act will end up issuing a blanket prohibition because it would hamper any development of the technology in the EU. However, it would be still possible to ban just certain use cases (such as in the case of biometric recognition) particularly prone to malicious exploitation.

For instance, an example could be the creation of deep fakes with pornographic content[44] which may be highly detrimental to a person's dignity or representing politicians with the intent of damaging reputations or destabilizing a geographical area (I am referring for instance to the deepfake of premier Zelensky declaring the defeat of

[40] European Parliament, Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, *Draft Report on amendments 310-538*, in Interinstitutional file 2021/0106(COD), PE731.563v01-00, 2022.

[41] A. C. Engler – A. Renda, *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act***,** cit., 22.

[42] In the words of the Council: «A new Article 52a and the related new Recital 70a have been added to clarify that general purpose AI systems should not be considered as having an intended purpose within the meaning of this Regulation. The new provisions also make it clear that the placing on the market, putting into service or use of a general purpose AI system should not trigger any of the requirements under the AIA». See Council of the European Union, *Presidency compromise text 8115/20*, in Interinstitutional file 2021/0106(COD), November 2022.

[43] Recital 15 of the AI Act.

[44] D. Harris, *False Pornography Is Here and the Law Cannot Protect You*, in *Duke Law & Technology Review*, 2019, 99 ss.

Ukraine at the wake of the Russian invasion)[45].

In fact, the legislator may mandate under general purpose AI systems requirements the inclusions of certain use cases to be filtered out by the safety components of the systems.

In the middle between the two extremes (no regulation and prohibition), risk regulation is a regulatory toolbox with its "policy baggage" that includes different policy options, ranging from design mandates to liability and conditional licensing[46]. In particular, «high-risk AI systems should only be placed on the Union market or put into service if they comply with certain mandatory requirements»[47].

Firstly, I think that it would be useful to further distinguish between the use (or misuse) of general purpose AI systems and their use to build downstream high-risk applications (e.g. a language model is fine-tuned to make medical assessment and utilized in an emergency room to identify priority codes).

a) General purpose AI systems in downstream high-risk applications

With regard to the latter, in consideration of the diffusion and popularity of these models (most of which are available open source), they may be utilized in downstream high-risk applications. In fact, after general purpose AI systems are made available, they typically require retraining and fine-tuning in order to be intended for use on a specific task. In most cases this operation likely amounts to a significant alteration *as per* art. 28, shifting the responsibility to the company that finetunes the general-purpose AI system, which then becomes the provider of a high-risk AI system, in case of a high-risk application[48].

However, downstream systems present their own set of difficulties: how could they comply with both the substantial and the documentation high-risk requirements? If upstream general purpose AI systems models were not regulated, it would be very complicated, nay impossible, to guarantee, for instance the quality of the dataset[49] when the model has been pre trained by a different company. The same objection may be raised in relation to the technical documentation required by the AI Act[50].

As regards the promotion of innovation through the valorization of the value chain and in response to this issue, both the Council and the Parliament impose cooperation requirements to support downstream applications' compliance with high-risk

[45] Interestingly, a model was created with the specific purpose of distinguishing between genuine and fake videos of President Zelensky. See M. Boháček - Farid H., *Protecting President Zelenskyy Against Deep Fakes*, in *arxiv.org*, 2022.

[46] M. E. Kaminsky, *Regulating the Risks of AI*, in *Boston University Law Review*, 2023, 103.

[47] Recital 27, which follows: «Those requirements should ensure that high-risk AI systems available in the Union or whose output is otherwise used in the Union do not pose unacceptable risks to important Union public interests as recognised and protected by Union law. AI systems identified as high-risk should be limited to those that have a significant harmful impact on the health, safety and fundamental rights of persons in the Union and such limitation minimises any potential restriction to international trade, if any».

[48] A. C. Engler – A. Renda, *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*, cit., 18.

[49] Art. 10 of the AI Act.

[50] As the documentation commonly available by general purpose AI systems providers do not cover all information required by Annex IV.

requirements.

The measure supports innovation and the utilization of these technologies and tries to avoid bottlenecks in which a few big incumbents squeeze new actors - especially small and medium enterprises (so called "SMEs") and startups which are particularly burdened by compliance costs - out of the market. Notably in this regard, art. 28a of the Parliament's compromise regulates unfair contractual terms unilaterally imposed on an SME or startup.

The Parliament in particular compels former providers to provide technical documentation and relevant information to the new provider for fulfilling regulatory obligations, while also considering third party suppliers and the protection of trade secrets through appropriate technical and organizational measures[51]. For instance, regarding foundation models provided as a service (such as through API access), recital 60f of the Parliament's compromise states that the cooperation with downstream providers should extend throughout the service, in order to enable appropriate risk mitigation, unless the training model and appropriate information[52] are transferred.

The attention towards the valorization of the value chain shows that both institutions want to sustain the drive towards innovation in the field of general purpose AI systems while concurring that society may benefit from a regulation mitigating possible risks arising from this new technology.

b) The regulation of general purpose AI systems

It is firstly necessary to individuate the scope of a possible regulation specific to general purpose AI systems.

In particular, a possible policy choice is that of considering possible risks arising in not high-risk applications as not severe enough - in the tradeoff between innovation and precaution - as to merit regulation *per se*. In this case, general purpose AI systems may only be regulated in case of possible high-risk applications. However, what if - as it is already happening - an LLM is utilized by a user in a high-risk sector? For instance, a Bolivian judge[53] utilized Chat GPT as one of the tools to assess the outcome of a case[54].

General purpose AI systems providers could limit their models so that they could not

[51] Art. 28, para. 5. Moreover, the Commission will create non-binding model contractual terms to assist high-risk AI system providers and third-party suppliers in drafting agreements that balance rights and obligations. These terms will be publicly available on the AI Office's website. The Parliament also proposes (art. 28a) protections for SMs and startups against unilaterally imposed unfair conditions. On balancing collaboration and disclosure, see also P. Hacker - A. Engel - M. Mauer, *Regulating Chat GPT and other large generative AI models*, in Proceedings: 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, 10-11.

[52] Namely, «extensive and appropriate information on the datasets and the development process of the system or restricts the service, such as the API access, in such a way that the downstream provider is able to fully comply with this Regulation without further support from the original provider of the foundation model», recital 60f of the Parliament's position.

[53] L.Taylor, *Colombian judge says he used ChatGPT in ruling*, in *theguardian.com*, February 2023.

[54] Annex III pt. 8 includes among the high-risk applications: «AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts». On the use of AI system in the Brazilian judiciary, see E. Villa Coimbra Campos, *Artificial Intelligence, the Brazilian judiciary and some conundrums*, in *sciencespo.fr,* March 2023.

be utilized in high-risk applications. However, such a solution has proved not to be robust as foundation models can be distracted by a change of context and prompt engineering[55]. Moreover, it is crucial not to create loopholes, e.g. allowing providers to simply state in their terms and conditions that their systems should not be used as a professional tool in certain sectors.

Notably, in high-risk applications, providers should also state and account for probable misuses. However, the requirement would not be applicable as the systems would not be identified as high risk and, in any case, it would be impractical if not impossible to impose compliance for every possible high-risk sector application in view of the possible (probable?) misuse.

However, high-risk applications aside, is there an inherent risk in the deployment of general purpose AI systems? *Id est*, standalone general purpose AI systems applications could be considered worth *per se* of legislative attention?

Indeed, they certainly present possible harms in relation to individual natural or legal persons (in fact, several lawsuits are being levied for defamation of copyright infringements). However, their peculiarity is the potential damage brought forward by the aggregate effect (and potential for propagation) of biased or hallucination-induced (but still plausible) outputs.

In fact, risks related to the propagation of biases and disinformation, and the overall resilience of the democratic system are more societal in nature. This interest towards more societal risks resembles the risk assessment[56] and mitigation[57] provisions the Digital Services Act (hereinafter "DSA") provides for very large online platforms and search engines (with over 45 million users), which regards systemic risks on illegal content; fundamental rights; civic discourse, electoral processes, and public security; gender-based violence, the protection of public health, minors and serious negative consequences to a person's physical and mental well-being[58].

Moreover, most of the high risk applications in Annex III derive from the possibility of a misleading output affecting (or replacing) a decision (*e.g. by a judge or a first responder) with significant effects towards a natural or legal person, whereas when we think about generative models like those used within Chat GPT or DALL-E, the user can be both professional and non-professional.*

Once identified the scope of the general purpose AI systems regulation, it is also necessary to determine what requirements they should comply with. Indeed, a number of difficulties may stem from the possibility of directly applying high-risk requirements (Title III Chapter 2 of the AI Act). Indeed, I think that it would be unfeasible to apply as is many (if not all) requirements for high-risk systems because they are purpose

[55] Prompt engineering is a relatively new discipline that studies the optimization of prompts (i.e. the query from the user of a generative AI model) to achieve more pertinent and efficient outputs.

[56] Art. 34 of 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).

[57] Ivi, art. 35.

[58] C. Djeffal, *Is the DSA Revolutionizing Algorithmic Risk Governance?*, in *Heinrich Boll Stiftung*, November 2022. Interestingly both Google and Bing (which has incorporated GPT 3.5 in its search engine) have been recently classified by the European Commission as very large search engines.

specific[59]. Therefore, if the AI system does not have a specific purpose and may be applied in different sectors, it would be very difficult to define, for instance, the representativeness of the datasets or the accuracy of the metrics as they are relational parameters (i.e. in relation to what task would they be representative or accurate?). The same consideration applies, for instance, to the risk management system and the instructions for the human oversight by the user as they would become very broad and un-useful for practical application. Furthermore, the risk management system (coupled with the newly proposed fundamental rights impact assessment[60]) should account for all possible risks (for health, safety and fundamental rights[61]) for all high-risk applications compatible with the system.

The most notable common feature between the Council's and the Parliament's proposal is probably the fact that both approaches propose watered down requirements based on the ones provided for high-risk requirements. Notably, the parliament has proposed further *ad hoc* requirements tailored for generative AI systems.

As illustrated in depth in the next sections, the Council proposes to regulate general purpose AI systems only when utilized in high risk systems (albeit with specific requirements). Conversely, the Parliament proposes a specific risk tier (adapted from the high-risk tier) among the AI Act's risk classification system.

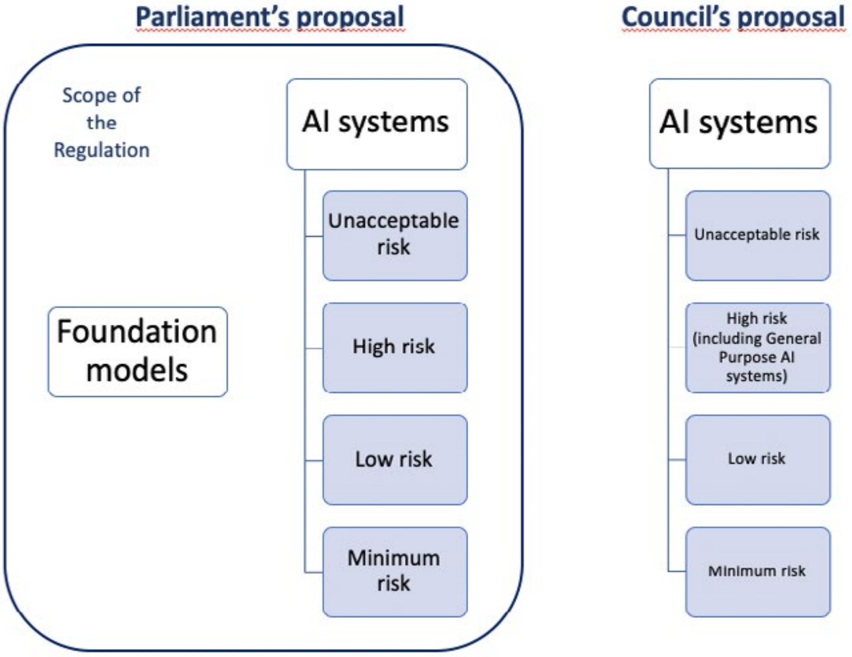


Figure 2 - A visual representation of how the proposed regulation by the Parliament (left) and by the Council (right) for General Purpose AI systems integrates within the AI Act architecture. Own elaboration.

[59] G. De Minico, *Too many rules or zero rules for the ChatGPT?*, in *BioLaw Journal*, 2, 2023, 493-494.

[60] The impact assessment was proposed at art. 29a of the Parliament's compromise.

[61] Art. 9 of the AI Act.

4. The Council's general approach

The Council compromise adheres to the idea of only regulating general purpose AI systems (art. 4b) if used as high risk AI systems (or components thereof). The Council tried to sidestep the negative effects of an as is application of the high-risk requirements, as the requirements of Title III Chapter 2 (i.e. requirements on the design and development of the systems)[62] will be tailored by implementing acts of the European Commission to account for general purpose AI systems' peculiarities[63].

The utilization of implementing acts is not a new legislative tool in the AI Act: it allows flexibility in the regulation as the Commission will be able to fine-tune and update the specific requirements for general purpose AI systems in the light of the latest state of the art. On the other hand, firstly, this timeframe seems surpassed by events, as the Council compromise was finalized in December 2022, before the general purpose AI systems hype and wide diffusion.

In fact, even if the legislative process was incredibly smooth, the regulation would only be published in the Official Journal in mid 2024, and a transitional period would be necessary before the entry into force. Therefore, the implementing acts detailing general purpose AI systems requirements would not be due earlier than December 2025.

Secondly, even if a tighter timeline was achieved, the delayed publication would not ensure legal certainty for general purpose AI systems and possibly impair innovation in the field in Europe. Moreover, it is pivotal to avoid regulatory capture through transparency and stakeholder participation[64]. In fact, although the requirements are technical in nature, the choice on what and how much to regulate is very much political: in the risk-based framework of the AIA, the choice on what requirements provide for general purpose AI systems is ultimately an assessment on the risk they pose to fundamental rights.

Moreover, providers can explicitly exclude all high-risk uses only if the exclusion is made in good faith and there are not sufficient reasons to consider that the model may be misused (art. 4c). In fact, especially for large language models[65], it would be quite difficult to exclude in good faith any use in high risk applications as also suggested by

[62] For example, provisions regarding risk management system, data governance, technical documentation and transparency instructions, human oversight an accuracy, robustness and cybersecurity. Notably, aside from the more substantive requirements of Title III Chapter 2, general purpose AI systems would not be subject to all obligations put forward by Title III chapter 3, regarding other obligations for providers as they would only need to comply with the following (art. 4b, para. 2): providing their name and trademark (art. 16 aa); conducting a conformity assessment (art. 16 e); registration (art. 16 f); corrective actions (art. 16 g); CE marking (art. 16 i); demonstrate conformity (art. 16 j); appointing an authorized representative (art. 25); EU declaration of conformity (art. 48); post market monitoring (art. 61); and sharing information with incoming competitors (art. 4b(5)).

[63] No later than 18 months after the publication (art. 4b).

[64] M. E. Kaminsky, *Regulating the Risks of AI,* cit., 79.

[65] In fact, Hacker et al. suggest that «Image or video models …generally count as high-risk systems», see P. Hacker - A. Engel - M. Mauer, *Regulating Chat GPT and other large generative AI models*, cit., 5.

Hacker et al.[66] and Engler and Renda[67].

As mentioned, the Council general approach also imposes a duty to cooperate with downstream providers (e.g. transmitting relevant documentation and information) in the case other providers utilize the models to create high-risk downstream AI applications. These provisions (artt. 4a to 4c) aim to strike a balance between burdening general purpose AI systems' providers with obligations not directly pertaining to their AI system and encouraging SMEs to integrate general purpose AI systems in their product.

Notably, requirements and obligations relating to general purpose AI systems do not apply to micro, small or medium enterprises (art. 55a, para. 3).

5. The Parliament's position: foundation models and general purpose AI systems

The two proposals also differ regarding the object of the regulation. Namely, the European Parliament proposes to differentiate between foundation models and general-purpose AI systems, while the Council only regulated the latter. The proposed definitions can be found in the table below.

	Council	Parliament
General Purpose AI system	*'general purpose AI system' means 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 (art. 3, pt. 1b).*	*'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 (art. 3, pt. 1d).*
Foundation model	/	*'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 (art. 3, pt. 1c).*

Table 1 - Definitions of general purpose AI and foundation models. Own elaboration.

[66] Ivi, 8.

[67] A. C. Engler – A. Renda, Reconciling the AI Value Chain with the EU's Artificial Intelligence Act, cit., 20 «Given the many categories of AI systems in products and standalone AI systems that can fall into the high- risk category of the AI Act, functionally this means that all general purpose AI systems would trigger these requirements».

The distinction between the two is mainly the training data (as foundation models are trained «on broad data at scale») and the possible use of general purpose AI systems for unintended purposes. Thus, foundation models include generative AI systems (e.g. Stable Diffusion or Chat GPT), which are defined by the Parliament 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 ("generative AI")»[68]. This choice (if confirmed by the trialogue) may lead to some confusion in the future application of the AI Act as the difference between general purpose AI systems and foundation models is not clear-cut. In fact, it mainly relies on the fact that only the foundation models are pre-trained and therefore ready to use. This definitory issue may seem trivial, but it is not as, under the Parliament's proposal, general purpose AI systems and foundation models, while similar in nature, will be regulated by two distinct regulatory frameworks. Indeed, even though art. 28b is collocated within Title III (pertaining to high risk AI systems), its regulation is conceptually separated: all foundation models shall be subject to art. 28b obligations, regardless of their purpose or risk level.

The justification for the different regulatory framework is not explicit in the text but could be found in the fact that foundation models are ready to be built upon to create new applications. Furthermore, many of these applications (such as Chat GPT) are directly accessed and utilized not only by professional users but also by end users. Therefore, also the risks mentioned above on the propagation of disinformation and biases would affect them directly.

Providers of foundation models are subject to the obligations of art. 28b when they make them available on the market or put them into service. Foundation models can ben «standalone model or embedded in an AI system or a product, or provided under free and open source licenses, as a service, as well as other distribution channels»[69]. It follows that Open AI is a provider both for businesses directly accessing its GPT-3.5 or 4 model via API and for natural or legal persons sending queries through Chat GPT. The same goes for other image or language models.

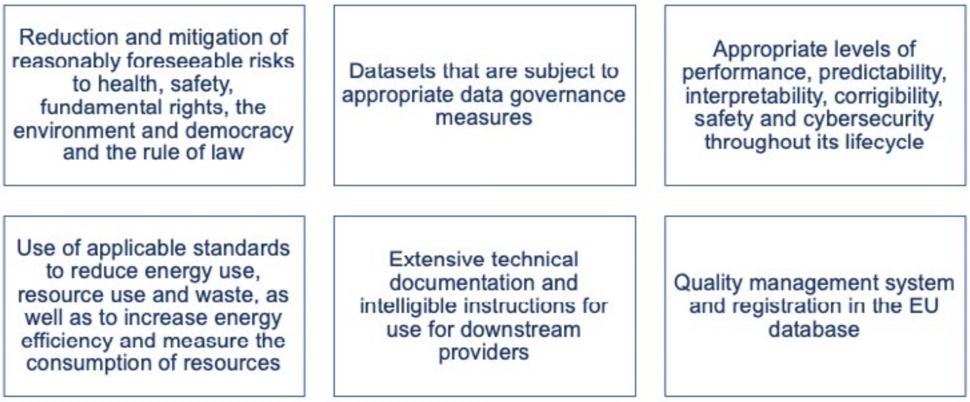
It should be noted that also the definition of «putting into service» provided by art. 3[70], which refers to the intended purpose of the system, should be updated accordingly as to allow its use with reference to foundation models.

According to art. 28b of the Parliament's position, foundation model providers shall ensure that the model is compliant with certain requirements listed in the figure below.

[68] Art. 28b, para. 4 of the Parliament's position.

[69] Art. 28b, para. 1 of the Parliament's position.

[70] «Putting into service' means the supply of an AI system for first use directly to the deployer or for own use on the Union market for its intended purpose» art. 3, pt. 11.



Reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law

Datasets that are subject to appropriate data governance measures

Appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity throughout its lifecycle

Use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency and measure the consumption of resources

Extensive technical documentation and intelligible instructions for use for downstream providers

Quality management system and registration in the EU database

Figure 3 - Visual representation of the requirements for foundation models proposed by the Parliament (art. 28b). Own elaboration.

Most notably, art. 28b provides for a risk mitigation obligation, widening the scope of possible risks. Indeed, high-risk AI systems shall identify and mitigate risks for safety, health and fundamental rights, whereas foundation models should also look out for «environment and democracy and the rule of law»[71] related risks, which, as mentioned resembles the DSA-related systemic risks.

In comparison with the risk management system (art. 9) set out by the AI Act for high-risk AI systems, art. 28b does not provide for monitoring obligations, therefore the risk management system only regards risks assessed «prior and throughout the development». This is highly inconsistent with a technology that - as discussed - continues to demonstrate emergent abilities and whose risks and limitations are not fully researched. Moreover, the models themselves are evolving, requiring maintenance and monitoring after their updates.

Furthermore, another critical aspect is the provision of technical documentation and instructions for use only for downstream providers. Indeed, if foundation models are accessible and can also be directly utilized by laymen, it would be consistent with foundation models' known specific risks to provide users with mandatory clear and appropriate information on the model's capabilities and limitations. The lack of clear and actionable instructions coupled with the lack of mandatory human oversight (as for high risk systems) may also enhance the so-called automation bias or *effet moutonnier*[72] as users may refer uncritically to the model's output.

The European Parliament also provided for cooperation requirements as to enable downstream high-risk applications (art. 28, para. 2; more on this in the next section) as former providers are obliged to provide any documentation, technical access and additional support required for the fulfillment of the obligations of the new (downstream) provider.

With particular consideration to Generative AI, para. 4 establishes that their providers shall:

- «comply with transparency provisions in art. 52;
- ensure adequate safeguards against the generation of content in breach of Union

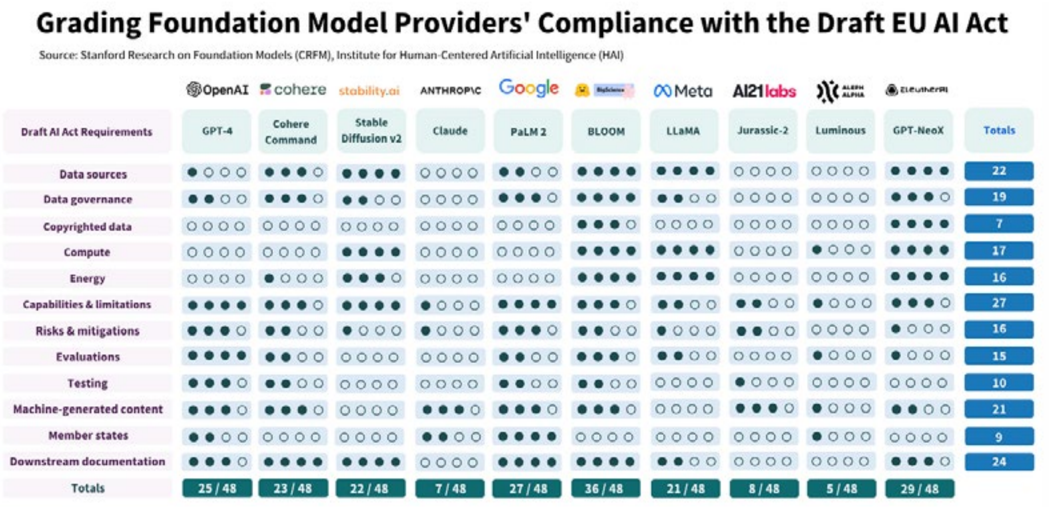
[71] Art.28b of the Parliament's position.

[72] A. Garapon - J. Lassègue, *Justice digitale. Révolution graphique et rupture anthropologique*, Paris, 2018, 239.

law in line with the generally acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression;

- document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law»[73].

A study by Stanford University[74], based on publicly available information, compared ten major foundation models providers on twelve (out of 22) selected Parliament-proposed requirements evaluated on a scale from 0 to 4. The image below shows the results of the research, with a breakdown of the grades awarded for every foundation model provider on every requirement analyzed.



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google		Meta	AI21labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Figure 4 – Evaluation of different foundation models on AI Act requirements. Source: Bommasani et al., Do Foundation Model Providers Comply with the Draft EU AI Act?, cit.

Notably, the highest-ranking requirement (and the only one regarding which all models were awarded at least one point) is "capabilities and limitations", showing a certain industry-wide attention to the topic; the worst-ranking requirement is linked to the publicity on the utilization of copyrighted materials.

The results show that currently, foundation models comply unevenly with AI Act's requirements; however, the study argues that the provisions proposed by the Parliament may lead towards «substantial progress towards more transparency and accountability».

The study also confirmed that, even though openly released foundation models score well on disclosure requirements, they perform worse on deployment control.

[73] Art. 28b of the Parliament's position.

[74] R. Bommasani - K. Klyman - D. Zhang - P. Liang, Do Foundation Model Providers Comply with the Draft EU AI Act?, in *Stanford.edu*, 2023.

5.1 The value chain

The Parliament's position also impacted the obligations of the different actors in the AI market. This section illustrates the applicable legal framework and indicates possible improvements.

The table below shows the obligations for each actor in the general purpose AI systems and foundation models value chain:

- foundation models are subject to the obligations of art. 28b when their providers make them available on the market or put them into service;
- general purpose AI systems are not regulated *per se*. They are (like all other AI systems) bound by the «General principles applicable to all AI systems» (art. 4a), but only indirectly, as the principles apply to other provisions such as Title III requirements for high-risk applications, the Code of Conduct or the harmonized standards but they do not create «new obligations under this Regulation»;
- only deployers of high-risk AI systems have oversight and monitoring obligations (art. 29); and
- downstream applications (who make a substantial modification to the AI system) from both general purpose AI systems and foundation models[75] may become high-risk AI systems.

The term deployer[76] in the Parliament compromise has substitute the term user of the original proposal (which could, in fact, have been misleading). Affected persons were not included in the table as they only have rights[77] and not obligations under the regulation.

Upstream AI system provider		Deployer	Downstream provider	Deployer (of the downstream provider)
General purpose AI systems	/	/	*High-risk: Title III Chapter 2 requirements apply*	*Obligations under art. 29*
	/	/	*Not high risk: Art. 52 may be applicable*	*Art. 52 may be applicable*

[75] When «*directly integrated into an high-risk AI system*» art. 28, pt. 2, of the Parliament's position.

[76] For example «any natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non- professional activity» art. 3, pt. 4. In order to quell any ambiguity, the Parliament also introduced (art. 3, pt. 8a) the notion of "affected person", i.e. «any natural person or group of persons who are subject to or otherwise affected by an AI system».

[77] Regarding high-risk AI systems, affected persons have the right to an explanation (art. 68c); information to be subject to the use of the high-risk AI system when it makes (or assist in making) decisions pertaining to natural persons (art. 29, para 6a). Affected persons also have a right to be informed about interacting with an AI and on the nature of AI-generated media (art. 52), regardless of the high-risk status of the AI system.

Foundation models	*Art. 28b requirements (referred to above)*	*Art. 52 may be applicable*	*High-risk: Title III Chapter 2 requirements apply*	*Obligations under art. 29*
			Not high-risk: Art. 52 may be applicable	*Art. 52 may be applicable*

Table 2 - Obligations of actors in the value-chain of general purpose AI systems and foundation models. Own elaboration.

This regulatory framework is disputable because its risk-based reasons are not apparent. Indeed, it is not clear in the text of the amended Regulation (articles or recitals) what risk assessment lead to the setting up of specific obligations (and that specific obligations) only for foundation models and not general-purpose AI[78] and to obligations (or lack thereof) for the different actors of the value chain.

Furthermore, even if providers of general purpose AI systems are not directly subject to documentation obligations, they have to collaborate with (and provide documentation and technical access to) their downstream providers; therefore also general purpose AI systems providers will also have to indirectly comply with many of the documentation requirements of the AI Act.

Moreover, the distinction between deployer and downstream provider warrants more attention.

a) Downstream providers

A new downstream provider is considered one under art. 28 when:

1. it places its name or trademark on a high-risk AI system (para. 1, pt. a);
2. it makes a substantial modification to a high-risk AI system which remains a high-risk AI system (pt. b); and
3. it makes a substantial modification to a not high-risk AI system (including a general purpose AI system) which then becomes a high-risk AI system (pt. ba).

In these cases, the former provider is not considered anymore the provider responsible for that system and shall provide technical documentation, access and assistance to the new provider for compliance purposes. Art. 28, para. 2 states that this applies to foundation models «directly integrated into an high-risk AI system»: An example could be a company integrating a foundation model via API as a chatbot evaluating candidates in a recruitment process.

However, it should be more clearly stated in the text that the implementation of a foundation model into a high risk system generates a new provider and a new AI system. Indeed, as currently worded, the equivalence may only refer to the mandated

[78] In fact, because of poor wording of amendment 34 (art. 28, para. 1, pt. ba) cited above mentions «a substantial modification to an AI system, including a general purpose AI system, which has not been classified as high-risk» one may conclude that general purpose AI systems may be classified as high-risk. However, pt. b), which pertains to «substantial modification to a high-risk system» does not mention general purpose AI systems at all.

cooperation[79].

Furthermore, it is not yet clear up to which point the documentation and requirements set out for foundation models (and the cooperation requirements for not high-risk general purpose AI systems) are specific enough for downstream providers having to comply (and attest compliance) with high-risk requirements.

b) Deployers

The term deployer indicates[80] the professional natural or legal person that utilizes the AI system «under its authority» and is subject to a number of obligations under art. 29 in case of high-risk AI systems. For example, this means that someone utilizing a language model to create a poem for a birthday card is not a deployer, whereas a business utilizing a generative AI system to create some images for a commercial presentation falls under this category.

Arguably, general purpose AI systems cannot have deployers as they are not ready for use and, if they were trained and applied to a specific high-risk application, the obligations for the provider would shift to the entity providing the training.

On the contrary, foundation models (which are pre trained, in the Parliament's definition) may very well have deployers. Notably, the only requirement for deployers is the one originally set out (albeit improved upon by both the Parliament and the Council) by the Commission's proposal, as art. 29 clearly only refers to deployers of high-risk AI systems. In fact, art. 52 sets out some transparency obligations applicable to the output of generative AI systems: *e.g.* disclosure requirements for systems interacting with natural persons (e.g. chatbots) or creating/manipulating media (e.g. deep fakes[81]). In particular, in the latter case, users (a relic from the change from users to deployers) «shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated».

Indeed, a foundation model deployer which utilizes a generative model should disclose the artificial origin of the output but deployers of high-risk systems have a large number of other obligations. That's because the deployer is the closest actor to the actual application of the system and therefore is the most qualified to e.g. ensure the representativeness of the dataset (art. 29, para. 3) and carry out data protection impact assessments (para. 6).

This legislative vacuum is inconsistent with the risks tied to the propagation of biases and disinformation mentioned above. Indeed, if these risks are such as to warrant an *ad hoc* regulation, more attention could be directed towards helping (as already mentioned) both laymen and professional deployers to utilize these models with awareness,

[79] «This paragraph shall also apply to providers of foundation models as defined in Article 3 when the foundation model is directly integrated in a high-risk AI system» art. 28, para. 2, of the Parliament's position.

[80] «'Deployer' means any natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity» art. 3, pt. 4 of the Parliament's position.

[81] Described in art. 52 of the AI Act as «text, audio or visual content that would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do, without their consent».

by providing them with adequate information on their functioning and limitations.
In particular, if a deployer utilizes foundation models in high risk applications (e.g. for considering the eligibility of a person for public benefits or for assessing and grading students' exams) it could be appropriate to impose some obligations of high-risk deployers such as human oversight and monitoring along with, in relation to systems referred to in Annex III, informing natural persons subject to a decision made (or assisted by) a high risk AI system[82].

6. Concluding remarks

In conclusion, the utilization of foundation models and general purpose AI systems brings forth, along with benefits, potential risks and harms. These risks are amplified by the aggregate effect and potential propagation of biased or hallucination-induced outputs, which may have societal implications and impact the resilience of democratic systems.
However, the Commission's proposal does not regulate general purpose AI systems nor foundation models. While the Parliament's position seems more promising, more focus on the specific risks related to the different actors in the value chain (including the provision of obligations for deployers, i.e. professional users) would be required. Moreover, when considering the requirements for general purpose AI systems and foundation models, it is crucial to determine the scope of such systems and the possible risks.
For instance, as regards foundation models, which can be directly accessed and utilized by citizens, it is appropriate to provide mandatory and clear information on the capabilities and limitations of these models and provide an iterative risk management system.
Overall, taking into consideration the potential risks and the need for clear regulations and guidelines, the final AI Act text will need to approach the deployment of these technologies with utmost care to protect the interests of individuals and society at large.

[82] Deployers of high-risk AI systems should ensure compliance with the instructions for use (also with the adoption of technical and organizational measures), human oversight, monitoring and maintenance of the appropriate robustness and cybersecurity measures. Most notably, the Parliament included also an obligation to inform natural persons that they are subject to a high-risk AI system utilized to make decisions (or assist in the decision-making process) (para. 6a).
Deployers shall also: ensure the use of relevant and representative data (if they have control over the input data), inform providers and relevant authorities in case of any serious incident or malfunctioning, carry out a data protection impact assessment and publish a summary thereof.