T4F Series
PAPER N. 11
a.y. 2023/2024

Foundation Models: Ethical and Regulatory Complications

LAURENCE BONAT,
DAVIDE CAVICCHINI, GIOVANNI VALER

Trento BioLaw Selected Student Papers

Jean Monnet EUROPEAN CENTRE UNIVERSITY OF TRENTO
Co-funded by the Erasmus+ Programme of the European Union
UNIVERSITÀ DI TRENTO Facoltà di Giurisprudenza

The paper was selected at the end of the course *Law and Ethics in Artificial Intelligence* a.y. 2023-2024, organized within the Jean Monnet Chair "T4F – TrAIning 4 Future. Artificial Intelligence and EU Law", coordinated at the University of Trento by Professor Carlo Casonato.

Foundation Models: Ethical and Regulatory Complications

Laurence Bonat, Davide Cavicchini, Giovanni Valer[*]

ABSTRACT: Foundation models, powerful AI models trained on vast datasets, offer immense potential for progress across various fields. However, their capabilities raise significant ethical questions regarding bias, alignment with human values, and the responsible use of their outputs. We examine these challenges and explore three regulatory approaches aimed at mitigating them: a horizontal approach with standardized rules for all models, self-regulation by industry players, and a tiered approach tailored to different risk levels. We argue that the tiered approach offers the most effective and sustainable solution, allowing for flexibility, efficiency, and continued innovation, while preventing harm by adhering to the risk-based approach of the AI Act.

KEYWORDS: Foundation Models; AI Act; Bias; Alignment; Ethical AI.

1. Introduction

Media coverage of the debate over the EU AI Rulebook approval skyrocketed in recent weeks. Negotiations came to an abrupt halt because some countries (namely France, Germany, and Italy) opposed any binding obligation for foundation model providers, a position not accepted by the European Parliament.

1.1. Foundation Models

As a new technology, there is no universally shared definition of foundation model. The term was originally proposed by Stanford University in 2021: «A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) 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»[1].
Foundation models have been introduced in the AI Act with Amendment 99 in June 2023: «Foundation models are a recent development, in which AI models are developed from algorithms designed to optimize for generality and versatility of output. Those models are often trained on a broad range of data sources and large amounts of data to accomplish a wide range of downstream tasks, including some for which they were not specifically developed and trained. The foundation model can be unimodal or multimodal, trained through various methods such as supervised learning or reinforced learning. AI systems with specific

[*] *Students of the University of Trento, Department of Information Engineering and Computer Science. Emails: laurence.bonat@studenti.unitn.it, davide.cavicchini@studenti.unitn.it, giovanni.valer@studenti.unitn.it.*

[1] R. BOMMASANI ET AL., *On the opportunities and risks of foundation models*, 2021, available at the link https://crfm.stanford.edu/report.html (last accessed 29/10/2024), p. 3.

intended purpose or general purpose AI systems can be an implementation of a foundation model, which means that each foundation model can be reused in countless downstream AI or general-purpose AI systems. These models hold growing importance to many downstream applications and systems»[2].
We can notice how, for the AI Act, foundation models are *often* trained on large amounts of data, while this is actually their salient feature. Moreover, there are other concepts with overlapping definitions (e.g., *General Purpose AI*), but we are not discussing such an issue.

1.2. Recent Developments

But why does the EU want to regulate foundation models? And why do some countries oppose further regulation? Foundation models acquire various capabilities in a wide range of tasks despite not being trained explicitly to do many of those tasks[3]. These *emergent* properties are rarely foreseeable and can cause unpredictable risks. Moreover, foundation models have led to an unprecedented level of homogenization: almost all state-of-the-art models (across a wide range of modalities) are adapted from one of a few foundation models. This is a heavy liability, as all AI systems might inherit the same problematic biases of a few foundation models. In particular, it is not possible to fully assess the harms of a foundation model before knowing how it will be deployed[4].
Aiming to address potential societal risks posed by the widespread use of foundation models, the European Parliament advocated for horizontal regulatory measures encompassing all models. However, some countries opposed, moved by the pressures of important companies fearing strict regulations (e.g., *Mistral AI* in France and *Aleph Alpha* in Germany), and pushed for mandatory self-regulation, in the form of codes of conduct. This situation prompted the European Commission to propose a tiered approach that seemed to be an acceptable compromise: the top tier of "very capable" foundation models, like GPT-4, would be subject to ex-ante vetting and risk mitigation measures. An extended description of the different approaches is provided in Chapter 3. Negotiations are still underway at the time of writing, with the potential failure to reach an agreement threatening the AI Act's approval.

2. The Importance of Data

To better understand how foundation models can pose threats to our society and require specific regulation, we need to take a step back and look at what differentiates them from other AI systems. As the Stanford definition points out, foundation models are not new in terms of architectures and technologies: what makes them different is the tremendous amount of data used for training. For this reason, we should pay particular attention to the data used for building foundation models. Complications regarding training data range from copyright violations to security and privacy issues, among many others; in this work, we focus our attention on the ethical impacts that data has on fairness and inequity, and then proceed

[2] 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*, 2023, available at the link https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html (last accessed 11/12/2023).
[3] T.B. BROWN ET AL., *Language models are few-shot learners*, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
[4] R. BOMMASANI ET AL., *cited*, p. 9.

showing how data gathering can affect the models.

2.1. Biased Data

To show how wrong things can go, we hereby present an example of the impact of biased training data on AI systems.

2.1.1. Norman AI

Norman is a psychopathic AI that sees death and destruction in everything[5]. It is trained to perform image captioning using data from a violent subreddit. When Norman undergoes a Rorschach test (an old and famous psychological test), the results are disturbing: the psychological interpretation of its captions confirms Norman's psychotic behavior.

2.1.2. Biases and Harms

Norman is a simple but clear example of how biased data can have concerning consequences. More generally, we can observe intrinsic biases, i.e., latent properties that can lead to harm in downstream systems. The most widespread form of bias is the representational one: people can be mis-, under- or over-represented in training data (e.g., data regarding African Americans excluded in training data). Foundation models have wide-ranging societal consequences that are challenging to understand: they act as intermediary assets that are not directly deployed, but rather serve as a foundation that is further adapted to myriad applications[6]. Intrinsic biases lead to extrinsic harms, i.e., people experiencing personal-level representational harm or abuse, but also group-level performance disparities (e.g., systems performing poorly on text or speech in African American English). These group disparities can have increasingly severe consequences, as exclusion from certain products can lead to further marginalization, amplifying bias in the data used to train foundation models and perpetuating a vicious cycle of disadvantage.

The Generalization Dilemma

Such a self-reinforcing spiral of discrimination is a problem that is far from easy to solve. This is because of the basic underlying idea of machine learning: *generalization*. In fact, foundation models are progressively getting bigger in terms of parameters and training data, but unfortunately, larger models have been observed picking up more social biases (even amplifying training data biases), as a side effect of the growing memorization capacity. In particular, the proportion of stereotypical errors compared to anti-stereotypical ones grows with the model size[7].

[5] MIT, *Norman AI*, available at the link http://norman-ai.mit.edu (last accessed 11/12/2023).
[6] R. BOMMASANI ET AL., *cited*, p. 131.
[7] Y. TAL, I. MAGAR, R. SCHWARTZ, *Fewer errors, but more stereotypes? The effect of model size on gender bias*, in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2022.

Memorization and Training Data Extraction

The capability of large language models to memorize, other than generalize, has been observed and studied in recent years, and this poses a significant risk as sensitive personal information can be recovered through a training data extraction attack[8]. This is a major problem since it is becoming more and more common to publish foundation models that have been trained on private datasets. For such a reason, research in this field has mainly focused on mitigating privacy risks[9]. In a similar way, biases in fine-grained sub-populations of data (therefore without an abundance of data that allows generalization) may cause unexpected behaviors and fairness concerns, giving rise to the need for unlearning[10]. However, there is currently no standardized framework for addressing ethical considerations in machine unlearning, leading to potential inconsistencies in unlearning outcomes[11].

2.1.3. Alignment

Foundation models are characterized by emergent capabilities they are not explicitly optimized for (e.g., the ability to control GPT-3 via "prompting" was an emergent phenomenon of which only the barest glimpses were evident in the smaller GPT-2 model). The misalignment between the foundation model's training objective and the desired behavior makes them even less explainable and spurs the challenge of alignment to human values and desired goals[12]. Alignment requires continuous monitoring of AI systems to identify and address unintended consequences as they emerge; a central method used to fine-tune AI systems to align with human goals is *reinforcement learning from human feedback*[13]. Despite being the most popular technique, little research has been done to overcome its problems and limitations, such as misaligned humans and bad oversight, among others (for instance *data poisoning*, see Section 2.2.5)[14].
To conclude, alignment for foundation models remains a complex challenge due to several factors: the inherent uncertainty and complexity of emergent capabilities, the limitations of current alignment methods, and the trade-off between performance and alignment. In the following, we have a glance at what can cause the biases we talked about, mainly focusing on the very first stages of data gathering.

2.2. Data Gathering and Quality of Data

After seeing the main ethical implications that foundation models pose, let's look at the process of how

[8] S. ISHIHARA, *Training data extraction from pre-trained language models: A survey*, in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 2023.
[9] J. JANG ET AL., *Knowledge unlearning for mitigating privacy risks in language models*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1, 2023.
[10] S. KUMAR ET AL., *Language generation models can cause harm: So what can we do about it? An actionable survey*, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023.
[11] T. SHAIK ET AL., *Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy*, 2023, available at the link https://arxiv.org/abs/2305.06360 (last accessed 31/10/2024).
[12] T. SHEN ET AL., *Large language model alignment: A survey*, 2023, available at the link https://arxiv.org/abs/2309.15025 (last accessed 31/10/2024).
[13] Y. WANG ET AL., *Aligning large language models with human: A survey*, 2023, available at the link https://arxiv.org/abs/2307.12966 (last accessed 31/10/2024).
[14] S. CASPER ET AL., *Open problems and fundamental limitations of reinforcement learning from human feedback*, in *Transaction of Machine Learning Research*, 12, 2023.

data is gathered and how the quality of this data can affect models.

2.2.1. AI Act Legislation

The importance of these foundation models and the data they use can be seen in the fact that one of the first articles about high-risk systems tries to regulate some obligations on how data is treated. Article 10 specifies how data should be collected and what processes a high-risk AI system should pursue to have good practices in labeling and collecting data. This article also specifies the methods that should be used when collecting and using the data for training purposes but remains quite vague about all of these concepts. It is also mentioned that it is important to have data that is "free of errors", and that sensible data should be treated according to the GDPR and with the correct data governance.

2.2.2. Data Gathering Process

After looking briefly at what the AI Act says about the collection of data, we will look at the actual state-of-the-art of data gathering. The vast number of parameters in large models requires vast amounts of information to learn and refine their abilities. This insatiable appetite for data necessitates readily available sources, and the internet provides a perfect solution. Nowadays large datasets are present on the internet. ImageNet is one of the most famous in the computer vision landscape and for LLMs is "Common Crawl". Much of the content collected for these datasets is scraped from the web, so questions about the privacy and copyright of this content arise. This method of collection does not have an explicit filter on the gathered content and could have the side-effect of having data that should not be considered because of its explicit content. For example, material from 4chan, a widely known forum for its non-filtered content, is contained in the common crawl dataset, and it's still even present in some cleaner-derived datasets like Google C4[15]. For example, GPT-3 used many different datasets, including "Common Crawl", for the training but prioritized higher quality datasets for the training step, using less time for the data from crawled and less refined datasets[16]. Collecting a great amount of data is quite challenging, and as we will see later, it can also create new problems with unbalanced data. Usually, image datasets tend to be cleaner, since they need to be labeled manually, but that is only sometimes the case. Now that generative AI is much more diffused we are seeing generated content flooding the internet and this could be detrimental to the creation of new AIs, as we will see in Section 2.3.

OpenAI Whisper

An example of a model that uses an amount of diverse data is OpenAI's Whisper, an automatic speech recognition system that was trained on a big labeled dataset (680,000 hours of content). This data is gathered from the web and contains different samples, from dirty audio with music in the background to clean audio samples. This approach led to major robustness with respect to smaller models because it

[15] K. QUACH, *Google's C4 ML training data drew from 4chan, racist sources*, 2023, available at https://www.theregister.com/2023/04/20/google_c4_data_nasty_sources (last accessed 11/12/2023).
[16] T.B. BROWN, *cited.*

created examples from a real-world scenario. A problem that could arise with this model is the fact that two-thirds of the training data is in English[17]. This heavily favors certain languages and could lead to potentially problematic behaviors, which we will explore in Section 2.2.4.

2.2.3. Emergent Abilities in LLMs

When we look at how AI models are trained, lots of data is used. In the last years, new abilities (in foundation models) have been discovered by increasing the number of parameters and the training data that is fed into the deep learning algorithm. Many tests performed on GPT-3 using different benchmarks demonstrate how new behaviors emerge from how many parameters are used, and thus how many training samples are available. While the scaling of parameters plays a crucial role in the development of emergent abilities in large language models, it's not the sole factor. Evidence suggests that models with fewer parameters can outperform big LLMs like GPT-3 in specific tasks where GPT-3 struggles. This phenomenon could be attributed to several factors, like superior training data, architectural differences, and data saturation. Scaling doesn't always guarantee better performance. Reaching a point where additional data provides minimal improvement or even hurts performance. Furthermore, studies have shown a correlation between the increasing size of language models and their tendency to generate toxic content. Considering approaches that reduce the harmful data in the training data is fundamental to avoid this kind of behavior[18].

2.2.4. Unbalanced Data

The world is full of diverse languages, each enriching our understanding and experience. Yet, when it comes to the development of artificial intelligence, a stark imbalance exists. The vast majority of data used to train AI models is in English, leaving languages spoken by millions around the globe under-represented. This imbalance between various languages, known as "the resourcedness gap", poses a significant challenge to the development of truly inclusive and multilingual AI. Multilingual language models are designed so that gaps between higher- and lower-resource languages are inferred. In such a way we can try to fill in the gaps of missing data, but this has side-effects. This is because many times the context matters, and some words do not correspond in different languages. For example, a multilingual model might associate the word "dove" in all languages with "peace" even though the Basque word for dove ("uso") can be an insult. The performance of AI models heavily relies on the amount of training data available in each language. This creates a "data divide" where languages with more data (e.g., English) enjoy better model performance, while languages with less data (e.g., minority languages) suffer from limitations in accuracy and functionality. Many times lower-resource languages stumble upon problems in the automatic filters used by big companies in filtering explicit or sensible content, because words may be interpreted in the wrong way. Proposed solutions in this regard include creating smaller research teams that try to help expand the material available in lower-resourced languages, by labeling and gathering more data. Even if these solutions are implemented correctly, there are still lots of problems that have to be taken into account.

[17] OpenAI, *Introducing Whisper*, https://openai.com/research/whisper (last accessed 11/12/2023).
[18] J. Wei et al., *Emergent Abilities of Large Language Models*, in *Transactions on Machine Learning Research*, 8, 2022.

Discrimination and the right to free speech are always in discussion when these instruments are misused[19].

2.2.5. Data Poisoning

In the realm of data gathering for generative AI, a contemporary battle has emerged in the realm of generative AI, particularly within the domain of AI-generated images and diffusion models. These models, falling under the category of foundation models, leverage large datasets and can be tailored for specific tasks with minimal data. Artists have begun employing techniques to actively poison the data[20] in a way that is invisible to the human eye but derails the ML models that we currently use. This emphasizes the need to consider public opinion and ethical considerations in the development and deployment of foundation models, as underscored in discussions about LAWS (lethal autonomous weapons systems) and potential societal backlashes. This scenario also serves as an important reminder that despite the impressive capabilities of current AI, it lacks true intelligence. Instead, it relies on statistical theories to approximate distributions that maximize the generation of data. The question if we can trust these models then naturally arises. The need for ethical scrutiny and public discourse becomes imperative in navigating the evolving landscape of artificial intelligence.

2.3. Model Degeneration

In Section 2.2.5 we saw the dangers of data being poisoned by human artists to harm the AI models' performance. But these models are not safe even from their own kind. In this section, we will discuss the problem that arises from the use of synthetic data in the training of AI models with a focus on the dangers of AI-generated content being left loose on the internet.

2.3.1. Synthetic Data Dilemma

As seen in Section 2.2, foundation models benefit from a huge amount of training data. Considering the effectiveness of existing models, we might consider as beneficial to incorporate some of their outputs into our training data to augment and enrich our datasets. But as often happens in this field not everything that shines is gold. Research on this issue revealed that models relying on statistical methods may degenerate into representations that merely reflect the mean of the underlying distributions, whether in language, images, or predictions, getting to a point where the model produces complete garbage[21]. This degeneration could also come from the *anchoring effect*, the human in the loop might not be enough to overcome this issue, since the human will rarely go against the decisions of the machines in fear of possible liabilities that might arise.

---

[19] G. NICHOLAS, A. BHATIA, *Lost in Translation: Large Language Models in Non-English Content Analysis*, in *Center for Democracy & Technology*, 2023, available at https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.

[20] S. SHAN ET AL., *Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*, 2023, available at https://arxiv.org/html/2310.13828v2.

[21] I. SHUMAILOV ET AL., *The curse of recursion: Training on generated data makes models forget*, 2023, available at the link https://www.cl.cam.ac.uk/~is410/Papers/dementia_arxiv.pdf (last accessed 29/10/2024).

AI Content Loose on the Web

One might consider simply excluding synthetic data from the training process. The reality is a bit harsher though. Recalling from Section 2.2, the conventional practice of gathering training data involves scraping the web for diverse forms of knowledge. However, the advent of AI-generated data becoming prevalent on the web poses a significant risk. Unintentionally, the model may be trained on synthetic data published on the internet without clear labeling. This is clearly shown by the fact that Google's search engine for images is starting to give as first results AI-generated images[22].

3. Regulation

Now that the main challenges have been outlined, it is imperative to explore the diverse approaches aiming to regulate foundation models. This involves an examination of how these strategies address the identified issues and an exploration of the measures adopted by individual entities to mitigate these concerns. To facilitate understanding, we separately consider the three key problems: alignment, quality of data, and model outputs.

Alignment

One critical challenge revolves around ensuring artificial intelligence aligns with ethical principles, including bias. Ideally, AI systems should adhere to the principles outlined in Luciano Floridi's AI4people[23]. These principles include beneficence, non-maleficence, autonomy, justice, and explicability. Addressing alignment issues requires a careful consideration of how AI systems can align with these ethical principles.

Quality of Data

Another crucial aspect is determining how to handle and collect data for training purposes. The selection and curation of training data play a pivotal role in shaping the behavior and performance of AI models. Regulatory measures need to be examined to understand how different approaches propose to manage and oversee the collection of training data.

Model Outputs

The final category focuses on regulating the dissemination of generated data in real-world scenarios. Controlling how models output data and ensuring responsible use are key components of ethical AI practices. Understanding how different regulatory approaches address the challenges associated with model outputs is vital for establishing a framework that promotes ethical and responsible AI deployment.

---

[22] K.X. TEO, *An image of Israel Kamakawiwo'ole shows Google search still can't tell AI-generated pictures apart from genuine ones*, available at https://www.businessinsider.com/google-search-ai-generated-pictures-israel-kamakawiwoole-midjourney-reddit-2023-11 (last accessed 11/12/2023).

[23] L. FLORIDI ET AL., *AI4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*, in *Minds and Machine*, 28, 2018.

3.1. Horizontal Approach

To explore the use of a horizontal approach in overcoming the challenges associated with foundation models, we try to apply current proposals from the EU that tackle the problems with foundation models. As mentioned in Chapter 1, the European Parliament and the Member States have still to come to an agreement on how or even if to regulate foundation models. In this section, we assume that the position of the European Parliament will prevail and foundation models will have to subside to the requirements defined in article 28b. This approach is the easiest to create from a regulation standpoint since it requires the specification of a single set of rules that applies to all providers. Another point in favor of this approach is from an economical standpoint since regulating only the downstream applications and providers in Europe could otherwise create an unbalance in costs and liability that may bring Europe to be behind in this sector. But, using a monolithic approach also makes it less future-proof and maintainable. We would have to obligate all providers to change their process if we want to change the regulation, instead of only targeting the outliers.

Alignment

In the realm of alignment, the EU's proposed AI Act takes substantial measures to ensure that AI systems conform to ethical principles. Article 28b outlines key principles, including the reduction and mitigation of foreseeable risks, datasets subject to appropriate data governance, and maintaining appropriate levels of performance, predictability, interpretability, corrigibility, safety, and cybersecurity throughout the AI system's lifecycle. Furthermore, the AI Act introduces additional provisions for high-risk systems, as highlighted in Article 8. These provisions encompass data and data governance (Article 10), transparency and the provision of information to users (Article 13), human oversight (Article 14), and criteria related to accuracy, robustness, and cybersecurity (Article 15).

Quality of Data

In the domain of data quality, Article 28b of the AI Act emphasizes the significance of datasets subject to appropriate data governance. It also necessitates a detailed summary of the use of copyrighted training data, ensuring a safeguard against the generation of content in breach of Union law. High-risk AI systems, as delineated in Article 8, must comply with specific requirements for the design and development of AI systems, including those related to data and data governance (Article 10).

Model Outputs

About this problem, there is little regulation for the distribution and diffusion of AI-generated data. The only article from the AI Act that can relate to this problem is Article 52, with Transparency obligations for the providers of foundation models. This entails that a human looking at a website or searching on a search engine has the right to know whether the content is AI-generated or not. And so it should also be possible for a crawler to exclude them.

3.2. Self-Regulation

The current proposal of the European Council on the other hand does not talk specifically about foundation models, only about the use in high-risk applications of GPAIS (General Purpose AI Systems), allowing for self-regulation (in the form of *codes of conduct*) from the industry. Current internal discussions, instead, call for a "mandatory self-regulation" of foundation models, mainly backed up by Germany, Italy, and France[24]. Note that this applies only to providers of foundation models, the high-risk requirements will still be enforced for the deployers of high-risk systems using foundation models. This approach surely fosters progress and growth in this field but brings with it the looming fear of repeating the mistakes of the last technological revolution of the 1990s. For some, taking a market-free approach in the EU would be a trojan horse for the EU's competitiveness and democracy, replicating the errors of the last technological revolution, which eventually led to the creation of Big Tech and major issues in terms of data protection, illegal content online or online market competition. While others say that «One should not consider that if Europe missed certain turning points in technology, it is trailing the race. When it comes to artificial intelligence, the game is still open»[25].

Alignment

A glaring example of companies rushing to regulate their own technologies is Microsoft and OpenAI which outlined mechanisms to identify, using an impact assessment of the system at hand, quantify, using both automated and human-assisted methods, mitigate, with different layers going down to the actual model all the way up to the user application, and operate, using a well-defined deployment and operational readiness plan, their most powerful generative AI models[26].

Quality of Data

The state of the art in the industry for now is data scraped from the web as delineated in the section Data gathering, disregarding the intellectual property of the content creators. In recent years, many lawsuits have arisen regarding copyright infringement, prompting major companies such as OpenAI to implement measures to safeguard against it. However, it remains a gray area as to whether or not certain uses can be categorized as fair use. Meanwhile, other major players started to have a more artist-focused approach, like Stability AI which for its latest generative model for music generation uses a licensed dataset[27]

[24] EURACTIVE, *France, Germany, Italy push for 'mandatory self-regulation' for foundation models in EU's AI law*, 2023, available at the link https://www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law (last accessed 11/12/2023).

[25] EURACTIVE, *Behind France's stance against regulating powerful AI models*, 2023, available at the link https://www.euractiv.com/section/artificial-intelligence/news/behind-frances-stance-against-regulating-powerful-ai-models (last accessed 11/12/2023).

[26] MICROSOFT, *Overview of responsible AI practices for Azure OpenAI models*, available at the link https://learn.microsoft.com/en-us/legal/cognitive-services/openai/overview (last accessed 11/12/2023).

[27] STABILITY.AI, *Stable audio: Fast timing-conditioned latent audio diffusion*, available at the link https://stability.ai/research/stable-audio-efficient-timing-latent-diffusion (last accessed 11/12/2023).

Model Outputs

To regulate the diffusion of model outputs private companies moved quicker than European legislators, recognizing the dangers of having AI content loose on the web. The newest Amazon model for image generator, Amazon Titan, for example, uses an invisible watermark, designed to help reduce the spread of misinformation by providing a discreet mechanism to identify AI-generated images[28]. Looking at the landscape of text, instead, OpenAI, clearly states on its website: «The role of AI in formulating the content is clearly disclosed in a way that no reader could possibly miss, and that a typical reader would find sufficiently easy to understand»[29].

3.3. Tiered Approach

This approach is a sort of midway from the ones previously outlined. It was proposed by the EU Commission to accommodate the requests from both the Parliament and the Council. It proposes a similar approach to the general one of the AI Act: define incremental requirements for different levels of foundation models[30]. The proposed tiered approach recognizes the diverse capabilities and potential risks associated with different foundation models. It proposes a graduated set of requirements that increase in stringency with the model's complexity, capabilities, and potential societal impact. This allows for a more balanced and targeted regulatory framework that avoids stifling innovation while ensuring responsible development and deployment. From the newest developments, models trained with computing power exceeding a certain threshold will now automatically be categorized as "systemic". Additionally, a new annex will outline criteria empowering the AI Office to make qualitative designation decisions either ex officio or in response to a qualified alert from the scientific panel. Criteria include the number of business users and the model's parameters trying to regulate both powerful models and models in the hands of millions of people (e.g., Gemini Nano[31]), and can be updated based on technological developments. Transparency obligations will apply to all models, including publishing a sufficiently detailed summary of the training data «without prejudice of trade secrets»[32]. AI-generated content will have to be immediately recognizable. Each tier will be subject to a corresponding set of regulatory requirements that are tailored to the specific risks and challenges associated with that tier. Note that, as before, the high-risk requirements are still valid. While the tiered approach offers a promising way to regulate foundation models, it also presents certain challenges. Establishing a transparent and objective framework for classifying models into

[28] AWS, *Amazon Titan image generator, multimodal embeddings, and text models are now available in Amazon Bedrock*, available at the link https://aws.amazon.com/it/blogs/aws/amazon-titan-image-generator-multimodal-embeddings-and-text-models-are-now-available-in-amazon-bedrock (last accessed 11/12/2023).

[29] OPENAI, *Sharing & publication policy*, available at the link https://openai.com/policies/sharing-publication-policy (last accessed 11/12/2023).

[30] EURACTIVE, *AI Act: EU commission attempts to revive tiered approach shifting to general purpose AI*, 2023, available at the link https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-commission-attempts-to-revive-tiered-approach-shifting-to-general-purpose-ai (last accessed 11/12/2023).

[31] GOOGLE, *Gemini: A family of highly capable multimodal models*, available at the link https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (last accessed 11/12/2023).

[32] EURACTIVE, *AI Act: EU policymakers nail down rules on AI models, butt heads on law enforcement*, 2023, available at the link https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-policymakers-nail-down-rules-on-ai-models-butt-heads-on-law-enforcement (last accessed 11/12/2023).

different tiers is crucial for ensuring fairness and consistency in regulation. The fast-paced nature of AI development necessitates a dynamic regulatory framework that can be readily adapted to accommodate new technologies and evolving risks. Striking the right balance between encouraging innovation and preventing potential harm is essential for ensuring a responsible and sustainable development of foundation models. Despite these challenges, the tiered approach offers several potential benefits. This approach focuses resources on models posing the greatest risks, allowing for more efficient and effective regulatory oversight. By avoiding overly burdensome regulations on lower-risk models, this approach fosters continued innovation and development in the field. The tiered framework can be readily updated and adjusted to address new challenges and technologies as they emerge.

4. Conclusion

The emergence of foundation models presents both tremendous opportunities and significant challenges. While their versatility and potential to revolutionize various industries are undeniable, their ethical implications and potential for harm cannot be ignored.

A Preferable Regulation

In this work we showed how biased data can lead to unforeseeable harm to society, and explored the key challenges of alignment, quality of data, and model outputs. We also examined three different regulatory approaches aimed at addressing these challenges: the horizontal approach, self-regulation, and the tiered approach. While each approach presents unique advantages and disadvantages, we believe that the tiered approach offers the most effective and sustainable solution for the following reasons:

- Flexibility: it allows for a more flexible and adaptable regulatory framework, accommodating the diverse capabilities and risks of different foundation models.
- Efficiency: it focuses on the most high-risk models, ensuring that regulatory efforts are directed where they are most needed.
- Innovation: by avoiding overly restrictive regulations for low-risk models, the tiered approach encourages innovation and the development of beneficial AI applications.
- Future-proof: it acknowledges the rapidly evolving nature of AI technology and allows for easier adaptation to future developments.

Therefore, we advocate for the adoption of a tiered regulatory framework for foundation models. This approach can help to ensure the responsible development and deployment of these disruptive technologies while maximizing their potential for good. In fact, it would be possible to have stricter requirements for the most powerful models, which are arguably also the most profitable ones; this should guarantee companies to actually being able to comply with the stricter rules. To effectively classify foundation models under this proposed regulatory framework, we suggest considering three key measures: the number of users, the capabilities of models, and a societal impact assessment. The number of users reflects the scale of influence, while capabilities encompass factors like computational power, complexity, and functionality. The societal impact assessment evaluates the consequences of deployment in specific

domains, such as healthcare (akin to Babylon in the UK[33]) or customer service for a website.
However, it is important to acknowledge potential drawbacks. The proposed measures might burden smaller players more in terms of making assessments, potentially impeding their ability to compete. Additionally, some of the measures, such as the "capabilities of models", may need clearer definition. It is worth noting that the Stanford University Center for Research on Foundation Models reached a similar proposal, recommending a two-tiered scheme: in the generic tier, foundation models are subject to disclosure requirements that improve transparency at minimal marginal compliance cost for companies. And in the high-impact tier, which triggers once foundation models show significant impact in society, more strenuous requirements are imposed[34].

Future Work

Moving forward, it is crucial to engage in a comprehensive and collaborative effort to ensure the responsible development and deployment of foundation models. By working together, we can harness the power of these transformative technologies while mitigating the associated risks and fostering a thriving and ethical AI ecosystem[35]. This necessitates ongoing dialogue between stakeholders from various sectors, including policymakers, researchers, industry leaders, and civil society. By taking these steps, we can ensure that foundation models are used for good and that the benefits of this powerful technology are shared by all.

[33] EMED, *Babylon Health*, available at the link https://www.babylonhealth.com (last accessed 11/12/2023).
[34] R. BOMMASANI ET AL., *Towards compromise: A concrete two-tier proposal for foundation models in the EU AI Act*, 2023, available at the link https://crfm.stanford.edu/2023/12/01/ai-act-compromise.html (last accessed 29/10/2024).
[35] EURACTIVE, *A(I) deal at any cost: Will the EU buckle to big tech?*, 2023, available at the link https://www.euractiv.com/section/artificial-intelligence/opinion/ai-deal-at-any-cost-will-the-eu-buckle-to-big-tech (last accessed 11/12/2023).