7. The Governance of Generative AI: Three Conditions for Research and Policy[1]

Fabian Ferrari

Abstract: The increasing permeation of society by generative AI systems like ChatGPT has given rise to a pressing task that remains unresolved: the design of future-proof governance mechanisms that ensure democratic oversight over those AI systems. To establish and examine this oversight, it is essential that generative AI systems can be opened up for regulatory scrutiny. This chapter argues that there are three overarching dimensions to structure research and policy agendas about the governance of generative AI systems: analytical observability, public inspectability, and technical modifiability. Empirically, the chapter explicates those conditions with a focus on the EU's Artificial Intelligence Act (AI Act). Those three conditions act as benchmarks to help perceive generative AI systems as negotiable objects, rather than viewing them as inevitable forces.

Keywords: foundation models, generative AI systems, regulatory objects, AI Act, transparency obligations, observability

Introduction: Navigating the AI policy landscape

Across the globe, governments find themselves confronted with a pressing challenge: How to establish robust oversight structures for generative AI

1 This chapter is based on the groundwork of two journal articles that appeared in *Nature Machine Intelligence* (Ferrari et al. 2023a) and *New Media & Society* (Ferrari et al. 2023b).

Van Dijck, José, Karin van Es, Anne Helmond, and Fernando van der Vlist, eds. *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values*. Amsterdam: Amsterdam University Press, 2025
DOI 10.5117/9789048562718_CH07

systems such as OpenAI's ChatGPT or Google's Bard? Consider Italy, which, in response to concerns about violations of user data privacy, imposed a temporary ban on ChatGPT in early 2023 (Satariano 2023). Similarly, Canada's privacy commissioner initiated an investigation into OpenAI, citing similar privacy concerns (Fraser 2023). Other governments are taking steps to seize the perceived economic advantages of generative AI systems. The United Kingdom, for instance, established the dedicated Foundation Model Taskforce, generously funded with £900 million of taxpayer money. The UK government envisions that these systems could "potentially triple" national productivity growth rates (Department for Science, Innovation and Technology 2023). These examples show that the global landscape of policy and governance approaches spurred by the increasing sophistication of generative AI systems is rapidly evolving.

Nevertheless, without a clear conceptual framework to interpret these fleeting, short-term developments as expressions of broader, long-term conditions for democratic oversight, it is difficult to navigate the swiftly changing AI governance landscape. When I refer to "democratic oversight," I mean the active involvement of democratic institutions, such as regulatory bodies, parliamentary committees, and scientific institutions that employ experts in machine learning and data governance, in the formulation, implementation, and monitoring of checks and balances for generative AI systems. In some cases, this oversight necessitates an understanding of how existing regulatory structures, such as data protection laws, are enforced in the context of generative AI systems like ChatGPT. Yet, in other cases, assessing democratic oversight may require an examination of specialized audit organizations tasked with scrutinizing the material properties of generative AI systems.

Generative AI systems are defined by their capacity to find patterns of dependencies between elements (e.g., words) in training datasets to produce new outputs with some variations based on those patterns. Such new outputs could be text, video, images, or sound. Regardless of the type of output, the same computational logic applies: there are underlying training datasets (e.g., Hemingway novels), there is some sort of pattern recognition, and there are outputs with some variations (e.g., Hemingway-inspired travel stories), such as changed pixel distributions or rearranged text data. Amid corporate-driven hype triggered by marketing terms like "artificial general intelligence" or "superintelligence," the stakes for problematizing the real-world properties of generative AI systems are high. As those opaque systems infiltrate economic, political, and cultural interactions, it is crucial to trace, theorize, and reimagine their globally interconnected governance structures. Oversight is necessary to avoid a further concentration of economic and

cultural power in the hands of a few powerful generative AI providers, as well as the misuse of generative AI systems in ways that may undermine democratic values (e.g., misinformation or hate speech).

Against this backdrop, the question of the chapter is: How can generative AI systems be rendered governable? In other words, how can those complex and multilayered systems be opened up for regulatory scrutiny? The primary challenge in answering this question stems from the fact that most advanced generative AI systems, including ChatGPT, are proprietary systems and their constitutive elements are shrouded in secrecy, making the establishment of democratic oversight mechanisms significantly more challenging (also see Hummel, in this volume). For instance, OpenAI has not disclosed details about the training dataset it used—gathered from the internet—to train ChatGPT for conversational purposes. We only know some basic information, such as the fact that an early version of ChatGPT was trained on a vast dataset of 45 terabytes, equivalent to around 300 billion words. This dataset comprised publicly available data from sources like Wikipedia, as well as data obtained under third-party licenses. Crucially, those sources remain undisclosed by OpenAI, hindering regulatory efforts to trace the provenance of training data.

However, while transparency regarding these training datasets is crucial, this chapter argues that "AI transparency" by itself is an insufficient benchmark for democratic oversight. Rather than utilizing the typically underspecified and vague concept of "AI transparency" as the key anchor point in research and policy, this chapter proposes a nested structure of three more holistic oversight conditions: analytical observability, public inspectability, and technical modifiability. First, democratic oversight requires a systematic observation of generative AI systems. Second, it mandates ensuring access to the properties of these models, whether for external inspectors or the general public. Third, it demands the capacity to modify generative AI systems based on those inspections. However, it is essential to stress that these conditions are interdependent. It is only when they come together that they create a coherent normative framework for research and policy upon which regulators can act.

To develop this argument, the chapter proceeds as follows. First, it situates the study of generative AI systems within the context of science and technology studies (STS) research on regulating multilayered objects. Second, it explains the three abovementioned conditions for democratic oversight, using the EU's Artificial Intelligence Act (AI Act) as a case study. Third, the chapter discusses the relevance of these conditions to study AI's regulatory futures.

Regulatory objects in science and technology studies

The field of STS has played a pivotal role in scrutinizing the dynamics between constantly changing governance subjects and regulatory frameworks. In Fisher's perspective (2014, 163), a "regulatory object" is defined as something perceived by regulatory actors as the focal point for regulation. To qualify as a regulatory object, it must be "understood by regulatory actors as the 'thing' to be regulated" (ibid.). What is the thing to be regulated, and how to systematically observe it over time?

STS scholarship suggests that the answer to this question is not simple. It depends on how complicated and layered the properties of the regulatory object are and how much they keep changing over time. An example is the governance of high-frequency trading algorithms that are used in stock markets. Seyfert demonstrates in his analysis of the German High Frequency Trading Act that "the demarcation of a manipulative trading algorithm is only a derivative second step after objectifying the algorithm as a distinct object" (2021, 6). In this case, the trading algorithm needs to be meticulously distinguished from both the trading platform and the trading firm. Although these three governance entities are inherently interconnected, it is pivotal to differentiate them analytically. Without a clear specification of what precisely constitutes the regulatory object, it remains impossible to make it publicly inspectable or subject to technical modifications.

Another clear example of this complexity can be seen in the regulation of genetically modified organisms (GMOs), such as transgenic agricultural seeds. In his study on how those organisms become new governance objects, Lezaun follows the "administrative practices and detection instruments able to track GMOs throughout the food production system, from the farm to the table" (2006, 501). The governance of those complex organisms is structured by overarching "infrastructures of referentiality" (ibid., 505), which consist of two parts. First, there is bureaucratic nominalism, whereby an unambiguous label is given to the regulatory object to make it categorizable in bureaucratic processes. Second, there is the standardization of detection methods, which helps in identifying the regulatory object. For example, GMOs need to be separated from non-GMOs, both for finding them in bureaucratic databases and detecting them through on-the-ground regulatory authorities.

Bureaucratic nominalism and standardized detection methods are also highly relevant in the context of generative AI systems. How can (and how should) generative AI systems be defined in regulatory frameworks? How can their use be detected in a standardized way, and how can changing use cases be observed? Those questions signify the importance of coherent

and clear regulatory definitions and agreed-upon governance standards. If different regulatory authorities within the same jurisdiction have different interpretations of the regulatory object, it can seriously hinder oversight processes. Conversely, when there are substantial differences in how different jurisdictions—such as the EU and the US—understand the regulatory object, it limits the effectiveness of cross-border regulatory systems. A lack of clarity regarding the precise definition of the regulatory object, including its boundaries and limitations, hinders efforts to govern generative AI systems. Both for research and policy in this area, a granular understanding of generative AI systems as regulatory objects with distinguishable properties is crucial: material items that can be observed, accessed, and modified.

Nonetheless, for this argument to carry empirical weight, it must be developed vis-à-vis an actually existing regulatory framework; it cannot remain an abstract theoretical claim. In the next section, I introduce the EU AI Act as an empirical case study that helps to bring to life the three interconnected conditions for democratic oversight.

Case study: The EU's AI Act and three oversight conditions

The EU's Artificial Intelligence Act relies on a risk-based approach through which different AI technologies get categorized by their risk level. Some, like facial recognition software, are labeled "unacceptable risk," while others fall into "high-risk" and "limited risk" categories. Because it has not come into force yet, the status quo is that the same corporate actors that produce generative AI systems like ChatGPT are also setting border-crossing standards for safety guardrails to mitigate repercussions. Corporate actors do not only own the means of generative AI production, but also the means of generative AI oversight. Even though they themselves call for setting up new AI regulations, they have a vested interest in defining the regulatory rules and principles, including the EU's AI Act (Perrigo 2023).

To influence AI regulations according to their strategic interests, industry-dominating AI producers can leverage consumer pressure. For example, Sam Altman, the CEO of OpenAI, the company which owns ChatGPT, has raised the prospect of withdrawing from the European Union's Digital Single Market should the company find it impossible to adhere to the EU AI Act. As of January 2023, reports indicated that ChatGPT was being used by more than 100 million individuals daily (UBS 2023). In November 2023, OpenAI claimed that 92 percent of Fortune 500 companies use ChatGPT (Porter 2023). The substantial user base, "making it the fastest-growing

consumer application in history" (Hu 2023), affords OpenAI's significant influence, as EU policymakers are unlikely to want to be seen as obstructing AI innovation or technological progress. Just as the ride-hailing company Uber has set up consumer petitions aimed at regulators, municipalities, and federal governments in the pursuit of corporate lobbying, similar efforts are likely in the case of OpenAI.

As a counterpart to corporate oversight over generative AI systems, the remainder of this chapter examines the EU's AI Act on the basis of three mutually dependent conditions for effective democratic oversight: analytical observability, public inspectability, and technical modifiability. This exploration underscores the necessity for establishing enduring high-level conditions that can withstand the swiftly evolving AI industry landscape.

Observing generative AI systems

To make generative AI systems governable for democratic oversight as material entities, we must begin by elucidating their constitutive elements and their place within broader industry dynamics. Without a well-informed analysis of how parts of generative AI systems—such as large language models (LLMs)—fit within platform ecosystems, and how they relate to other entities (e.g., platform companies), we cannot distinctly delineate them as regulatory objects. As STS scholarship (Lezaun 2006) shows, a clear delineation of what needs to be governed according to precisely defined technical parameters and detection methods is crucial. Only after pinning down what, we can address how generative AI systems can be governed.

Crucially, a dynamic and processual perspective is required when dealing with ever-changing AI systems, rather than relying on static or rigid governance procedures. As Rieder and Hofmann convincingly argue, "unlike transparency, which nominally describes a state that may exist or not, observability emphasizes the conditions for the practice of observing in a given domain" (2020, 3). Consequently, the term observability is more appropriate than alternatives like "AI transparency" or "AI explainability" because it stresses how generative AI systems play a dual role: they form the foundation of new products and services, including chatbots and media creation tools, while relying on underlying computational infrastructure for their technological functioning. In other words, when analytically observing generative AI systems in the pursuit of governing them, it is insufficient to focus on one dimension, such as highly visible applications like ChatGPT. Rather, the crux is to acknowledge "generative AI" as a complex relationship, in which computational infrastructure, LLMs, and consumer-facing applications are intricately intertwined (figure 7.1).

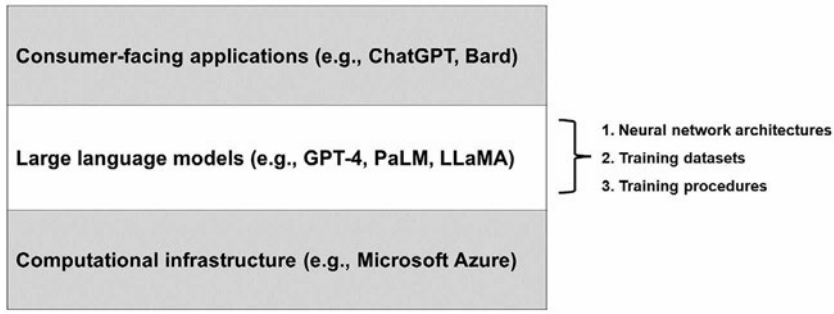

Figure 7.1. Observable dimensions in the context of generative AI systems.

One could interpret figure 7.1 through the "platformization tree" metaphor (Van Dijck 2021). At the top of this tree are the consumer-facing AI applications (e.g., chatbots), which depend on generative foundation models beneath them to run on a daily basis. These models serve as the central trunk of the tree and demand specific computational resources,[2] including graphics processing units (GPUs). They are not isolated lines of code; they exist within a broader economic and industry context. Therefore, understanding their features as regulatory objects necessitates an examination of the political-economic context in which they operate and reshape cultural practices. Alterations to the model as the middle layer of this tree can have ripple effects on both the upper and lower layers of the system. While consumer-facing applications often draw the most attention from policymakers and the public, the inner workings of the underlying models tend to be difficult to grasp. Similar to high-frequency trading algorithms, generative models are highly changeable and dynamic due to constant developer modifications and user interactions (e.g., training or fine-tuning).

The European Commission's initial proposal in April 2021 lacked explicit provisions for generative AI systems. However, this has since evolved, influenced by the introduction of ChatGPT. A pre-final version of the AI Act, disclosed by a European Parliament official in January 2024 (Caroli 2024), no longer categorizes generative AI systems as high risk. Instead, it includes specific provisions for providers of general-purpose AI models. As the text outlines, "these models are typically trained on large amounts of data, through various methods, such as self-supervised, unsupervised or reinforcement learning" (Caroli 2024, 48). Specific requirements include

2 When it comes to computational resources, there is a complex global network of actors that includes chipmakers like AMD and Nvidia, semiconductor firms like TSMC and Qualcomm, assemblers of server farms like Supermicro and Inventec, and data center providers like Equinix.

"disclosing that content was generated by AI," "preventing the model from generating illegal content," and "publishing summaries of copyrighted data used for training" (European Parliament 2023). However, the provisional agreement of the AI Act lacks technical specifics about what components of those models need to be made inspectable to regulators. Crucially, not all types of models are included as regulatory objects, only so-called general-purpose AI models. Indeed, high-profile commercial models such as Google's Pathways Language Model (PaLM) and OpenAI's generative pre-trained transformer (GPT) are increasingly crucial as gatekeeping tools at the center of the AI ecosystem.

To effectively observe generative AI systems, it is valuable to differentiate between three observable components of those models: neural network architectures, training datasets, and training procedures (figure 7.1). Public discussions tend to predominantly address training datasets, mainly due to evident concerns related to copyright and privacy issues. Nonetheless, it is crucial to acknowledge that both neural network architectures and training procedures are equally significant components for ensuring democratic oversight. In terms of neural network architectures, most generative foundation models rely on the transformer architecture, which was initially introduced by Google researchers in 2017 (Vaswani et al. 2017). Google's paper that introduced this architecture was publicly accessible. Given this openness, it subsequently served as a fundamental technical basis for OpenAI in the development of their own models that underlie ChatGPT. Understanding proprietary and—therefore closed—neural network architectures presents a greater challenge compared to their open-source counterparts. A similar complexity surrounds comprehending the training procedures, particularly when it comes to fine-tuning models for specific tasks like conversational use, which often remain inaccessible to regulatory scrutiny. Promoting "ethical" self-regulation among companies could hinder significantly democratic oversight because there is no economic incentive for these firms to make their foundation models transparent. Given the competitive nature of the AI industry, companies have a vested interest in maintaining opacity.

To push back against this intentional opacity, the next section builds on those insights to specify in more detail what layers of information need to be made accessible by developers of generative AI systems, and to whom. Observation alone is futile without regulatory access to key parameters of those quickly evolving systems. This mutability raises the issue of how regulators can gain insight into the inner workings of generative AI systems.

Inspecting generative AI systems

The condition of "public inspectability" in generative AI oversight raises the question of how those systems should be subjected to public scrutiny. In this context, "inspectability" means that generative AI systems are available for in-depth examination at the most detailed level. This idea of public inspectability presents a policy dilemma regarding whether generative AI systems should be entirely open, as advocated by open-source proponents, or entirely closed, as argued by their proprietors. However, this is not a binary question, and it is useful to allow for granular differences. Solaiman (2023) usefully introduces a gradient of inspectability levels, ranging from fully closed (whereby systems remain sealed off by their developers and inaccessible to the public) to fully open, where they are entirely accessible to the public. The granular levels of access in between include gradual or staged access, hosted access, cloud-based or application programming interface (API) access, and downloadable access.

The concept of public inspectability includes a delicate balance between public values like safety and security and ideals of openness and democratic control. Public inspectability also intertwines with concerns about detecting and managing misinformation, manipulation, and unauthorized use of resources. For example, an open-source model lacking adequate safety measures (e.g., Stanford's Alpaca model, which was taken offline due to safety concerns) may not be a better option than a closed, proprietary model that does have robust safety controls. So-called "model cards" have emerged as a standardization tool for AI developers to comprehensively document all key aspects of generative AI systems, including domain-specific training datasets, biases, and ethical considerations (Mitchell et al. 2019). In cases involving closed models, opacity sometimes masquerades as superficial transparency: model cards previously released by OpenAI and Meta attracted valid criticism from the research community (Birhane et al. 2021) and policymakers (Blumenthal and Hawley 2023) for being severely under-detailed, possibly intentionally so. Consequently, the concept of public inspectability prompts a challenging question: What components of proprietary generative AI systems should be made inspectable, to whom, and for what purposes?

Based on cross-disciplinary research conducted at the Governing the Digital Society focus area with Antal van den Bosch and José van Dijck, a structure of a five-layer model of different types of information about foundation models was developed (Ferrari et al. 2023b). Table 7.1 provides an overview of this basic structure, making a structural distinction between "types of information" (e.g., training datasets), "formats of information" (e.g.,

text), and "ways to access" this information. In the following paragraphs, this chapter elaborates on those types of information in the context of the EU AI Act. While appreciating the complex hierarchy of three observable dimensions of generative AI systems (figure 7.1), the focus of this chapter in answering this question is specifically on the middle (trunk) layer of the emerging "generative AI tree": generative foundation models. Although it is equally important to analytically dissect the components of consumer-facing AI applications as well as computational infrastructure, the remainder of this chapter focuses on the model dimension.

Table 7.1. Five Layers of Inspectable Properties of Generative Foundation Models

Type of information	Format of information	Access to information
1. Training datasets for foundation models	Text (or images, video, etc., depending on the model)	Model card, inspection of training datasets
2. Domain-specific training datasets for fine-tuning	Text, dialogue, text labels	Model card, inspection of fine-tuning dataset
3. Neural network architectures	Config file	Model card, inspection of architecture in config file
4. Trained models (with all trained parameters)	Weights file	Model card, inspection of parameters in weights file
5. Scripts for training and output generation	Computer code	Model card, inspection of scripts in model's code

Accessing training datasets for foundation models

In the relevant literature, the problem is widely acknowledged that the training datasets of many high-profile commercial algorithmic systems, whether designated as AI or not, remain uninspectable to external examination. A substantial body of research has grappled with the problem of algorithmic opacity (Brevini and Pasquale 2020). In the context of proprietary generative AI systems, developers tend not to give access to the datasets they have trained their models on, and at best give non-exact pointers to the datasets.

OpenAI's GPT models were trained on openly available data and data acquired under third-party licenses. GPT-3.5, for example, was trained on 45 terabytes of text data, which adds up to approximately 300 billion words extracted from public sources like Wikipedia, CommonCrawl, and GitHub, but also from undisclosed other sources. Open models, such as Meta's LLaMA, by contrast, tend to give out pointers to the training datasets, but often leave out technical details on selections and applied pre-processing methods.

Accessing domain-specific training datasets for fine-tuning
Commercial systems typically shield these datasets, citing competitive reasons. A counterexample is Google's Med-PaLM, where Google has been open about which databases were used for fine-tuning PaLM for medical purposes, including more than 200,000 question-and-answer sets from medical exams, and consumer questions and reference answers by the US National Institute of Health (Singhal et al. 2023). Open systems often refer to widely used benchmark data and evaluation scores of the systems themselves on these data.

Accessing neural network architectures, trained models, and training scripts
In the case of openly available foundation models, comprehensive information about neural network architectures is usually provided, often available on platforms like HuggingFace. For example, the BLOOM model was shared openly via the HuggingFace platform (Scao et al. 2023). However, in closed (commercial) systems, such details are not disclosed and may be underspecified in model cards. Regarding trained models and their parameters, open systems typically offer full access, providing the weights file and all necessary information about the configuration of the neural network architectures. Conversely, closed-source systems in the commercial domain usually do not provide complete access. When it comes to scripts for training and generating output, closed systems may offer code to interact with their APIs (i.e., without downloading the model), ensuring controlled access to the model. Open systems, on the other hand, often provide a range of scripts and code, which is frequently contributed by multiple users, thereby enabling collaboration and validity checks.

It is important to note that the pre-final version of the provisional agreement of the EU's AI Act lacks explicit details for conducting these audit processes (Mökander et al. 2023), as it does not sufficiently distinguish between the different levels of information and their formats mentioned earlier. As table 7.1 illustrates, different layers of information come in different formats, such as configuration files versus text or image data, necessitating distinct approaches for external inspection—for example, reviewing files versus examining datasets. Instead of merely lamenting the limits of algorithmic opacity, this chapter emphasizes the significance of identifying specific technical details that require examination. Simply using the term "AI transparency" without specifically defining which layers of information about generative AI systems should be inspectable and for whom is therefore inadequate. As the following section illustrates, this

detailed understanding is also crucial for determining how the properties of generative AI systems can be technically modified.

Modifying generative AI systems

The third and final oversight condition, technical modifiability, poses the question: How can and should the material properties of generative AI systems be modified through regulatory action, and what are the reasons for doing so? The term "modifiability" here refers to making basic or fundamental changes to a regulatory object to shape it according to specific public values. Therefore, this condition explores how proposed regulatory frameworks, such as the EU's AI Act, may reshape the material properties of generative AI systems.

In science and technology studies scholarship, the condition of technical modifiability can be grounded in Jasanoff's (2004) concept of co-production, which posits that "knowledge and its material embodiments are at once products of social work and constitutive of forms of social life." The way in which co-production works in practice depends on the motivations for modifying regulatory objects through regulating them in the first place. For example, in the case of chemicals, safety issues may prevail. As Fisher puts it, "the role of co-production may be recognized in relation to the question of the *safety of a chemical* but not much the identity of the chemical itself" (2014, 165, emphasis added). In the case of high-frequency trading algorithms, regulatory measures to modify those algorithms are driven by concerns about financial manipulation (Seyfert 2021). Similarly, in the context of generative AI systems, apprehensions regarding misinformation, manipulation, and unauthorized usage of sources (e.g., copyright infringement) may motivate regulatory actions.

Consider the policy goal of curbing the spread of "misinformation" by generative AI systems like ChatGPT. In this context, the study of how specific technical alterations to the system could achieve less misinformation becomes crucial, encompassing enhancements like more robust safety filters, digital watermarks, or more effective content moderation systems. Watermarking, as an AI governance tool, is not a speculative notion but a present regulatory practice in certain countries. China's Cyberspace Administration has implemented regulations that limit the production of AI-generated content lacking clear labels, stipulating that citizens must not use "technical means to delete, tamper with, or conceal relevant marks" (Edwards 2022). In this scenario, watermarking serves as a form of censorship. The identification and subsequent modification of AI technology that is categorized as potentially harmful to "the legitimate rights and

interests of the people" and detrimental to "national security and social stability," offer autocratic oversight regimes ample room for interpretation and enforcement of digital censorship. By defining the scope of generative AI systems that can become subject to technical modifications as widely as possible, regulators gain greater control over choosing which AI systems fall within the purview of restrictive regulatory frameworks.

In the European Union, by contrast, the modifiability of generative AI systems must be firmly anchored in democratic principles, encompassing public values such as openness, privacy, and autonomy. For example, while provisions mandating the use of watermarks can have relevance for democratic oversight within the EU legal framework, it is vital to prevent dominant firms like OpenAI or Google from having a monopolistic influence on determining the application of watermarking techniques. In line with the previously mentioned five layers of information pertaining to generative AI systems, I hold that each of those layers also offers distinct approaches and rationales for technical modifiability (see table 7.2).

Table 7.2. Five Layers of Modifiable Properties of Generative Foundation Models

Type of information	Modifiable by whom?	Rationale for modifying
1. Training datasets for foundation models	Model developers (in the process of pre-training)	Reduction of bias or harmful content in AI-generated outputs, enforcement of data protection regulations (e.g., GDPR)
2. Domain-specific training datasets for fine-tuning	Model deployers (in the process of fine-tuning)	Control over post-processing of the foundation model (e.g., ChatGPT's RLHF layer), enforcement of data protection regulations (e.g., GDPR)
3. Neural network architectures	Developers (pre-training), users (trainable models)	Control over (and reduction of) the size, training time, and energy consumption; retraining on selected training datasets
4. Trained models (with all trained parameters)	Developers (pre-training), users (trained from scratch)	No reason to modify
5. Scripts for training and output generation	Developers (pre-training), users (trained from scratch)	Control over replication, retraining from scratch and generation of output

Modifying training datasets for foundation models

In the case of pre-trained models, the possibility of modifying training datasets has already been concluded, and making significant modifications

to LLMs through retraining with adjusted datasets is challenging. End users lack the capability to alter the training data, even if they have access to it. The ability to determine and modify training datasets is exclusive to home-grown models that are pre-trained from scratch without any prior pre-training. In such cases, datasets can be either omitted from the training process or modified to mitigate bias issues or minimize the generation of harmful outputs. This specific form of debiasing, known as intrinsic debiasing, is a complex area of research (Orgad et al. 2022). Notably, a significant portion of intrinsic debiasing research has concentrated on gender debiasing, employing methods that mask or counterbalance gender-specific terms like gendered pronouns, first names, and other gender-specific language. Beyond bias reduction, compliance with data protection regulations such as the General Data Protection Regulation (GDPR) necessitates the ability to modify training datasets.

Modifying domain-specific datasets for fine-tuning
Users typically have the ability to adjust or fine-tune some aspects of pre-trained models to better suit their needs. This option may be open to the user for models that allow some sort of fine-tuning on top of the pre-trained foundation model. If so, domain-specific datasets should be entirely modifiable by users and regulators. However, if fine-tuning concerns proprietary data that is an integral part of the released model, it may be vital to have access to this dataset to be able to understand better which toxic, badly formed, and other unsuitable output is filtered away (and which is not) by this post-processing layer. The modifiability of this type of data is also crucial for the enforcement of data protection regulations (e.g., patient data covered by the GDPR to train domain-specific medical chatbots.)

Modifying neural network architectures, trained models, and training scripts
When looking at pre-trained models that are not hidden behind an API, their workings can often be packaged in a downloadable architecture config file that contains information like weights and how the model is structured. However, once a model is trained, these aspects are fixed and cannot be easily modified without potentially causing errors. Even if the model is fully open, the model weights are simply the end result of training procedures and modifying them manually makes no sense (as it will likely harm performance). The consumption of fewer energy resources, which is one of the requirements of the AI Act, could be attained by architectural modifications. However, this only makes sense at the stage of pre-training; it would be too late to implement modifications at a later stage. When it

comes to scripts for training and output generation, only some output-related code might be shipped along with a pre-trained foundation model (e.g., for fine-tuning), and modifying it is necessary for various downstream tasks. When training from scratch, scripts can be modified.

The EU AI Act includes the need to "train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law" (European Parliament 2023). However, precise technical details on how to enforce modifications of foundation models effectively remain unspecified, echoing the ambiguity of AI transparency obligations. Table 7.2 offers a structured pathway to delve into the technical adaptability of foundation models. Yet, addressing the challenges of technical modifications requires comparative studies on future compliance.

Conclusion

Analytical observability, public inspectability, and technical modifiability are best understood as normative benchmarks against which the actual empirical properties of oversight structures pertaining to generative AI systems can be measured in terms of democratic control. Those three conditions offer a practical roadmap for making generative AI systems negotiable in regulatory terms. For instance, even if these models are not fully open to the public, we can gauge their level of openness by resorting to the criterion of public inspectability. This real-world perspective counters the prevailing narrative that emphasizes long-term AI risks, often characterized by terms like "super-intelligence" or "artificial general intelligence"—notions that are often used in corporate efforts to influence policymakers, including those involved in shaping and negotiating the EU's AI Act as part of its trialogue (Perrigo 2023).

When finishing this chapter in December 2023, it was still unclear whether and how the EU's AI Act may come into being. A crucial topic of debate related to the inclusion (or exclusion) of foundation models and their providers in the AI Act. Germany and France, for example, suggested excluding those providers, which would mean that there are no specific obligations for inspectability or modifiability (Bertuzzi 2023). This exclusion would place a significant compliance burden on smaller EU companies using these models. Meanwhile, the owners of these models could avoid accountability. Only a few prominent foundation models, such as Google's PaLM, Anthropic's Claude, OpenAI's GPT-4, and Meta's LLaMA models, serve as the basis for various generative AI start-ups in the EU. Despite

claims of promoting AI democratization, the AI industry is dominated by a small number of platform monopolies. Since Microsoft and Amazon, as infrastructure providers, benefit from the widespread use of generative AI systems, they lack inherent economic motivation to prevent misuse by bad actors. Therefore, any regulatory efforts that focus solely on addressing issues like fake news without tackling the uneven power dynamics only offer a surface-level solution.

Generative AI systems should not be seen as escaping the grip of democratic control. Granted, their material complexities differ from other regulatory objects expounded upon in science and technology scholarship. Take the example of aircraft. Aircraft are fixed objects, comprising many components like engines, propellers, and other parts. Before they can enter the market, regulatory bodies must grant approval for all these components. Generative AI systems, on the other hand, consist of a small set of component types, essentially artificial neurons, but the multitude of connections between them allows for an immense variety of architectural configurations. This means that generative AI systems can have endless architectural shapes and use cases: they may influence elections, precipitate public scandals, and shape the norms of cultural production according to their probabilistic logic.

Regardless of how generative AI systems present themselves to public scrutiny in the future, oversight mechanisms need to be grounded in their material properties—not in speculative ideas about human extinction. If we perceive AI systems as carriers of existential risks, their right to exist precludes democratic negotiation. There is an urgent need to dispel this notion of AI systems as inescapable forces imposed upon society, instead recognizing them as observable, inspectable, and modifiable objects. In this way, democratic negotiations will become inescapable forces imposed upon generative AI systems.

References

Bertuzzi, Luca. 2023. "EU AI Act 'Cannot Turn away from Foundation Models,' Spain's State Secretary Says." *Euractiv*, November 17. https://www.euractiv.com/section/artificial-intelligence/interview/eu-ai-act-cannot-turn-away-from-foundation-models-spains-state-secretary-says/.

Birhane, Adeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv. https://arxiv.org/abs/2110.01963.

Blumenthal, Richard, and Josh Hawley. 2023. "Hawley and Blumenthal Demand Answers from Meta." Senator Josh Hawley, June 6. https://www.hawley.senate.gov/hawley-and-blumenthal-demand-answers-meta-warn-misuse-after-leak-metas-ai-model.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. "On the Opportunities and Risks of Foundation Models." arXiv. https://arxiv.org/abs/2108.07258.

Brevini, Benedetta, and Frank Pasquale. 2020. "Revisiting the Black Box Society by Rethinking the Political Economy of Big Data." *Big Data & Society* 7(2): 1–4. https://doi.org/10.1177/2053951720935146.

Caroli, Laura 2024. "AI Act Consolidated Version." LinkedIn. https://www.linkedin.com/posts/dr-laura-caroli-0a96a8a_ai-act-consolidated-version-activity-7155181240751374336-B3Ym.

Department for Science, Innovation and Technology. 2023. "Initial £100 Million for Expert Taskforce to Help UK Build and Adopt Next Generation of Safe AI." Gov.uk. https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai.

Edwards, Benji. 2022. "China Bans AI-Generated Media without Watermarks." *Ars Technica*, December 12. https://arstechnica.com/information-technology/2022/12/china-bans-ai-generated-media-without-watermarks/.

European Parliament. 2023. "MEPs Ready to Negotiate First-ever Rules for Safe and Transparent AI." European Parliament, July 14. https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai.

Ferrari, Fabian, José van Dijck, and Antal van den Bosch. 2023a. "Foundation Models and the Privatization of Public Knowledge." *Nature Machine Intelligence* 5: 818–20. https://www.nature.com/articles/s42256-023-00695-5.

Ferrari, Fabian, José van Dijck, and Antal van den Bosch. 2023b. "Observe, Inspect, Modify: Three Conditions for Generative AI Governance." *New Media & Society.* OnlineFirst. https://doi.org/10.1177/14614448231214811.

Fisher, Elizabeth. 2014. "Chemicals as Regulatory Objects." *Review of European, Comparative & International Environmental Law* 3(2): 163–71. https://doi.org/10.1111/reel.12081.

Fraser, David. 2023. "Federal Privacy Watchdog Probing OpenAI, ChatGPT Following Complaint." *CBC News*, April 4. https://www.cbc.ca/news/politics/privacy-commissioner-investigation-openai-chatgpt-1.6801296.

Hu, Krystal. 2023. "ChatGPT Sets Record for Fastest-Growing User Base: Analyst Note." Reuters, February 2. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Jasanoff, Sheila. 2004. *States of Knowledge: The Co-production of Science and the Social Order.* London: Routledge.

Lezaun, Javier. 2006. "Creating a New Object of Government: Making Genetically Modified Organisms Traceable." *Social Studies of Science* 36(4): 499–531. https://doi.org/10.1177/0306312706059461.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, 220–29. https://arxiv.org/abs/1810.03993.

Mökander, Jakob, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. "Auditing Large Language Models: A Three-Layered Approach." *AI and Ethics.* https://doi.org/10.1007/s43681-023-00289-2.

Orgad, Hadas, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. "How Gender Debiasing Affects Internal Model Representations, and Why It Matters." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2602–28. https://aclanthology.org/2022.naacl-main.188/.

Perrigo, Billy. 2023. "OpenAI Lobbied the EU to Water down AI Regulation." *Time*, June 20. https://time.com/6288245/openai-eu-lobbying-ai-act/.

Porter, Jon. 2023. "ChatGPT Continues to Be One of the Fastest-Growing Services Ever." *The Verge,* November 6. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference.

Rieder, Bernard, and Jeannette Hofmann. 2020. "Towards Platform Observability." *Internet Policy Review* 9(4): 1–28. https://policyreview.info/articles/analysis/towards-platform-observability.

Satariano, Adam. 2023. "ChatGPT Is Banned in Italy over Privacy Concerns." *New York Times*, March 31. https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html.

Scao, Teven Le, et al. 2023. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model." arXiv. https://arxiv.org/abs/2211.05100.

Seyfert, Robert. 2021. "Algorithms as Regulatory Objects." *Information, Communication & Society* 25(11): 1542–58. https://doi.org/10.1080/1369118X.2021.1874035.

Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620(7972): 172–80. https://www.nature.com/articles/s41586-023-06291-2.

Solaiman, Irene. 2023. "The Gradient of Generative AI Release: Methods and Considerations." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 111–22. https://doi.org/10.1145/3593013.3593981.

UBS. 2023. "Let's Chat about ChatGPT." https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html.

Van Dijck, José. 2021. "Seeing the Forest for the Trees: Visualizing Platformization and Its Governance." *New Media & Society* 23(9): 2801–19. https://doi.org/10.1177/1461444820940293.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30* (*NIPS 2017*), edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus and S. Vishwanathan and R. Garnett. Red Hook, NY: Curran.

About the Author

Fabian Ferrari is assistant professor in cultural AI at Utrecht University.