You shall understand what you regulate: A case study of the ANPD-Meta case in Brazil

João Pedro Padua[1]

[1]Faculdade de Direito – Universidade Federal Fluminense (UFF)
Rua Presidente Pedreira, 62, Ingá – 24.210-470 – Niterói – RJ – Brazil

joaopadua@id.uff.br

Abstract. *Regulating artificial intelligence is a primary need of current legal systems. However, bridging legal, policy and technical expertises to achieve good regulation is hard. We illustrate this difficulty by doing a short case study of a decision by the Brazilian* Autoridade Nacional de Proteção de Dados–ANPD *suspending Meta's collection of data to train its large language models. I try to demonstrate that the lack of technical knowledge led the ANPD to issue a bad decision, that ended up backfiring. I end by discussing implications for regulatory best practices.*

1. Introduction

Regulation of Artificial Intelligence (”AI”) is a pressing matter in legal systems and political debates today [Guha et al. 2023]. Harms associated with the use of AI and LLM models are commonly reported, as are whistleblower allegations of company greed and lack of regard for the public interest[Bond and Allyn 2021].Although it is critical to continue this debate and create smart regulation, there is also the danger that bad regulations address bogus problems, fail to deal with actual ones and even backfire, hindering progress.

In this paper, I present a short case study of an application of Brazilian legal rules by the *Autoridade Nacional de Proteção de Dados–ANPD* (National Authority on Data Protection), that, I will argue, can be categorized as an example of bad use of regulatory power. In Section 2, I summarize the case. In Section 3, I address how LLMs are trained and why this knowledge makes patent how the ANPD's arguments are not sound. Section 4 ends the paper with a brief discussion of potential pitfalls in AI regulations.

2. ANPD's decision main arguments

On July 7th, 2024, the ANPD published a decision to ”immediately suspend [Meta's] new privacy policy [...] in what regards the use of personal data to train generative AI systems”[Autoridade Nacional de Proteção de Dados 2024a]. ANPD is a federal agency created by Law 13.709/2018 (*”Lei Geral de Proteção de Dados–LGPD”* to oversee data privacy and data protection online.

ANPD's decision was based on an opinion by Director Miriam Wimmer [Wimmer 2024]. It followed the flagging by ANPD's staff of the then recent policy change by Meta, about using content publicly posted on all its platforms to train their generative AI (”gen AI”) products, especially the Llama family of LLMs[1]. According

[1]”Llama”is the brand name Meta gives to its LLMs, the same way OpenAI calls theirs ”ChatGPT”and Google/Alphabet does the same with their ”Gemini”models

to the opinion, Meta's policy change violated the LGPD in four aspects, only the first of which concerns the scope of this paper.

The opinion formulated this point as the "inadequate use of the legitimate interest legal hypothesis [according to Article 7, item IX, of the Law 13.709/2018]"[Wimmer 2024] and accounted for it thusly:

> [Meta's new privacy policy promoted] the treatment of sensitive personal data, the non-observance of the legitimate expectations of the owners, and the non-compliance with the principles of legitimate ends (*"princípio da finalidade"*) and of necessity.

Article 7 enumerates instances in which, by way of exception, user's personal data can be used by service providers for their own ends. Item IX allows for this use "when necessary to fulfill the legitimate interest of the controller [i.e., the service provider] [...]". ANPD, then, is denying that there is such a legitimate interest for Meta in training LLMs.

ANPD seems to be assuming that training LLMs works as training of other types of machine learning/AI algorithms, especially in the realm of discriminative AI–of which recommendation algorithms are examples [Meta 2023]. In these types of models, user's data are important as a whole, connected to the user, and categorized as a specific variable in the recommendation model (*ex.gr.*, number of clicks, time in a page, likes, age, gender, location, etc.). The assumption that users' data will be used as a whole is made clear in a part of the opinion where Director Wimmer quotes an ANPD's staff report:

> the indiscriminate treatment of photos, images, videos and audio recordings, especially through the use of artificial intelligence systems, can reveal political, religious, union affiliations and sexual preferences of its owners, which characterize them, immediately, as sensitive personal data [...].

This line of argument shows how the relevant legal framework for data protection can be misconstrued by a lack of understanding of the technical phenomenon being regulated.

3. How does LLM training actually work

Current LLMs take advantage of massive deep learning architectures, most notably the Transformer architecture [Vaswani et al. 2017], with billions or trillions of learnable parameters. LLM training algorithms are so-called *self-supervised*, which means that they are exposed to language data without tagging for features they are trying to predict. Instead, the objective is to assign probabilities for what the next token will be in a sequence that gets progressively unmasked during training [Naveed et al. 2023]. By training in trillions of words–and then subjecting them to additional instruction fine-tuning [Ouyang et al. 2022]– these models begin, apparently by brute force, to be able to mimic language use and linguistic interaction with human-level competence [Wei et al. 2022].

This training is preceded by the treatment of the language data via another algorithm called *tokenizer*. Current LLMs converge to a tokenizer algorithm called *byte-pair encoding* or BPE [Sennrich et al. 2016]. Meta's Llama models use a version of it as well, since the Llama 3 series [Team Llama @ Meta AI 2024]. BPE starts by breaking words

into characters and converting them to byte objects (to account for various alphabetic character systems in different languages). The algorithm, then, proceeds by finding the most common pair of bytes and merging them into a new object. By iterating this for long enough, BPE actually recreates whole words or word chunks. These tokens are what LLMs are trained on and what make up their vocabularies.

The conjunction of the tokenization of language data and the training algorithms makes the specific content of the words, phrases and sentences that comprise the dataset matter mostly because of their combinatorial properties, and less because of their specific content or reference. For example, if I use the same tokenizer as the Llama models[2] to preprocess a made-up phrase containing sensitive personal data, I get the list of tokens displayed in Table 1:

Tabela 1. Example tokenization of text containing sensitive data using BPE

Original sentence	**Tokenized version**
José da Silva Souza Monteiro, brasileiro, advogado, 43 anos, CPF 053.123.456-78	['José', ' da', ' Silva', ' Souza', ' Monte', 'iro', ',', ' brasileiro', ',', ' advogado', ',', ' ', '43', ' anos', ',', ' CPF', ' ', '053', '.', '123', '.', '456', '-', '78']

The resulting tokens do not always follow word boundaries, especially around sensitive information, like the tax code number or the name. Also, since the training algorithm of an LLM is trying to spot which tokens follow each other, and since personal information, by definition, does not occur frequently, an LLM will give a close to 0 probability that, say, token "053"will follow token "CPF", in the context window of tokens "José"and "Souza". In other words, for training LLMs, sensitive information is basically meaningless, unless one assumes data leakage before tokenization.

4. Discussion: Why regulators need to understand what they regulate

This short case study illustrates the pitfalls of trying to regulate complex AI phenomena from a perspective of legal doctrine combined with only a basic understanding of the phenomena one is trying to regulate. Because of the specific ways in which LLM models are trained, regulation needs to address them differently than it addresses other ML/AI models. More generally, before making regulations and construing and enforcing ones that already exist, authorities need to strive to better understand how different AI models work, lest they risk making rules that are not effective and leaving sensitive points without regulation [Guha et al. 2023].

Moreover, badly conceived rules and decisions may backfire. ANPD's decision probably did little to protect users of Meta products and services and Brazilian citizens, but it did incentivize Meta to halt the rolling out of gen AI services in their products in Brazil. At the time of the final version of this paper, Meta and the ANPD reached an agreement and the ANPD eventually reconsidered its first decision and allowed Meta to use data from their platforms to train their LLM models

[2]Since the Llama 3 series, Meta uses a slight modification of the Tiktoken algorithm originally created by OpenAI for training their models. For the toy demonstration in this paper, I used the tiktoken library with a Python code. The code and other materials for this paper can be accessed through GitHub: https://github.com/joaoppadua/meta_anpd.

[Autoridade Nacional de Proteção de Dados 2024b]. Therefore, in Brazil, the case had a sort of happy ending. In Europe, however, decisions similar to the first ANPD's by their data privacy authorities are cited by the company as the reason why they are still not making the gen AI tools in their platforms available to European citizens [Fried 2024].

In AI, even more than in other areas, it appears that regulation grounded in wrong premises will lead to outcomes just as bad as not regulating at all. Cooperation between the government, companies, and the Third Sector seems to be the key to avoiding this.

Referências

Autoridade Nacional de Proteção de Dados (2024a). Despacho Decisorio nº 20/2024/PR/ANPD.

Autoridade Nacional de Proteção de Dados (2024b). Despacho Decisorio nº 33/2024/PR/ANPD.

Bond, S. and Allyn, B. (2021). Facebook whistleblower tells Congress products hurt kids and weaken democracy.

Fried, I. (2024). Scoop: Meta won't offer future multimodal AI models in EU.

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Surani, F., Bommasani, R., Raji, I. D., Cuéllar, M.-F., Honigsberg, C., Liang, P., and Ho, D. E. (2023). The AI Regulatory Alignment Problem. Technical report, Stanford Human-Centered Artificial Intelligence.

Meta (2023). The AI behind unconnected content recommendations on Facebook and Instagram.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A Comprehensive Overview of Large Language Models. *Journal of Latex*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceeding of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin. ACL.

Team Llama @ Meta AI (2024). The Llama 3 Herd of Models. Technical report, Meta.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wimmer, M. (2024). PROCESSO Nº 00261.004509/2024-36, Voto n.º 11/2024/DIR-MW/CD.