Lara Dal Molin

Infrastructure Governance for Generative Machine Learning Systems

Introduction and Objectives

This position paper follows up on the presentation "Large Language Models as Gender-Shaping Infrastructures" delivered at the Governance by Infrastructure Workshop at the University of Lausanne. In the field of Science and Technology Studies (STS), infrastructure refers to both tangible and abstract entities necessary to human activities, including equipment, protocols and standards (Bowker et al., 2010). Throughout the second half of the 20th century, Artificial Intelligence (AI) and its applications – most prominently Machine Learning (ML) – infiltrated large sociotechnical systems and became integral parts of contemporary infrastructures (Weizenbaum, 1976; Russell and Norvig, 2010; Goodfellow, Bengio and Courville, 2016). This paper is concerned with specific AI technologies known as Generative Pretrained Transformers (GPTs). These are especially sophisticated instances of Language Models (LMs) – distributions of words and sentences in a language, such as English, that, when coupled with particular AI algorithms, generate human-like text (Jurafsky and Martin, 2008). In other words, GPTs are generative systems: as opposed to traditional ML systems, they do not simply produce labels or categories, but produce data (Harshvardhan et al., 2020; Suzuki and Matsuo, 2022). GPTs are integrated in numerous, far-reaching applications including machine translation, text summarisation and dialogue, an example of the latter being conversational assistants (Li, 2022). However, GPTs reportedly display stereotyped language, an instance of algorithmic bias that reflects real-world products of ideological hierarchies, such as gender-occupation associations (Barocas et al., 2019; Sheng et al., 2019). It may be argued that current conceptualisations of infrastructure, while they contemplate categorisation, do not consider the social implications of generative technologies. This paper presents, discusses and problematises current formalisations of infrastructure and infrastructure governance in relation to generative ML systems – particularly GPTs. Subsequently, it proposes hybrid, qualitative-quantitative ontologies and epistemologies as theoretical tools to update and rebalance the structural and infrastructural influence of GPTs.

Infrastructures, Categorisation, and Traditional Machine Learning Systems

In Sorting Things Out, Bowker and Star (1999, p.319) that "everyday categories are precisely those that have disappeared into infrastructure, into habit, into the taken for granted" and that "the moral questions arise when the categories of the powerful become the taken for granted; when policy decisions are layered into inaccessible technological structures; when one group's visibility comes at the expense of another's suffering". It may be argued that this standpoint appears to closely relate concepts of infrastructure and infrastructure governance to categorisation. Further, according to Bowker and Star (1999), infrastructures may be conceptualised as mechanism that uphold existing configurations of categories. The Social Worlds framework, formalised by Clarke and Star (2006, p.115), suggests that infrastructures may be considered instances of "frozen discourses", which manufacture and shape avenues between social worlds and larger structures. Particularly relevant for this analysis is Clarke and Star's (2006) selection of the adjective *frozen*, which emphasises rigidity as an attribute of infrastructures, and consequently their capability to uphold social structures. This particular configuration of infrastructure may be verified through classificatory ML systems. Commonly associated with traditional ML paradigms, classification systems execute some extent of decision-making based on real-world examples (Russell and Norvig, 2010; Goodfellow et al., 2016). Applications of ML classification systems are widespread and comprise, for instance, clinical diagnosis, image classification and email filtering, as well as more elaborate product recommendations and predictions of individual preferences (Nichols, Chan and Baker, 2019). Consistently, results of classification algorithms may be considered deterministic, as they are associated with pre-defined, standardised categories. The implications of such systems are especially problematic when associated to human personal characteristics, such as gender and ethnicity. In this case, it may be argued that the determinism that characterises such technologies translates to structural determinism, defined as the reinforcement of pre-existing, often oppressive, categorisations (Noble, 2018). Therefore, in the context of categorisation, infrastructure may be understood as an entity, or collection of entities, that enable social reality and Social Worlds to operate the way they do.

Generative Pretrained Transformer (GPT) Language Models

Recently, paradigms in ML shifted as a result of breakthroughs in AI algorithms and increasing computing power (Goodfellow et al., 2016). In the field of Natural Language Processing (NLP), this manifested in the integration

of AI algorithms, such as ANNs, and large-scale LMs (Min et al., 2021). It may be suggested that the generative capabilities of large LMs, including GPTs, mark a stark separation from previous, traditional ML systems. This poses a significant challenge to the fields of STS and Social Data Science (SDS) as, in the effort of interpreting the increasing amount of unorganised data that fills online spaces, these must account not only for human-generated material and its natural inequalities, disparities and controversies, but for machine-generated content too. In the foundational paper *On the Dangers of Stochastic Parrots: Can Language Models be too Big?*, Bender et al. (2020) consider several potential risks associated with GPTs. Prominently, the authors problematise stereotyped language in the context of personal characteristics, such as gender. While there can be different sources of bias in ML, stereotyped bias in GPTs has been specifically associated with problematic training data (Basta, Costa-jussà and Cosas, 2019; Kurita et al., 2019). Indeed, when datasets are assembled with the objective of scalability, hegemonic narratives are most likely to be retained (Bender et al., 2020). Therefore, generative ML technologies do not solely reinforce pre-existing categorisations, but manufacture novel narratives and discourse where superiority is associated with certain personal attributes over others. Consequently, it may be argued that GPT models disrupt the traditional Social World concept of infrastructures, understood as "frozen discourses", through an additional layer of complexity deriving from their generative and unpredictable functionalities (Clarke and Star, 2008). Therefore, questions arise regarding infrastructure governance, namely, how to update concepts of infrastructure governance accordingly, and rebalance the structural and infrastructural influence of generative technologies? And why is this important?

Prominently, in the forward-thinking article *The Computer for the 21st Century*, Weiser (1991, p.94) affirms that "the most profound technologies are those that disappear", and continues "they weave themselves into the fabric of everyday life until they are indistinguishable from it". It may be argued that this is applicable to GPT models and their applications. On one hand, the impact of these is far-reaching. For instance, through a radical feminist analysis of generative language technologies, Dillon (2020) problematises the ascription of feminine traits to speech-based conversational agents, such as Apple's (2022) Siri. This does not solely reinforce an association between femininity and labour, but also represent the outsourcing of historically feminine work to machines (Hester, 2017; Costa and Ribas, 2019; Dillon, 2020). On the other, the lack of transparency and interpretability of GPT models substantially convolutes the effort of implementing improvements. Indeed, due to their complex architecture, these artefacts are characterised by substantial opacity, and progress in this space is tentative and often poorly understood by software developers themselves (Carabantes, 2020). Concerning the black-boxed nature of some technological artefacts, Offenhuber (2017, p.15) suggests that "design practices share concern for what should be hidden or exposed". Additionally, Posner (2018) suggests that "partial eyesight", articulated in decentralised, modular systems, is a characteristic of modern technologies and one of the building blocks of the globalised economy. It may be argued that GPT models "disappear", to quote Weiser (1991, p.94), not solely for their extensive social implications, but also for their black-boxed, modular and decentralised architecture. However, it may be additionally suggested that the "disappearance" of generative technologies comes at a cost, which is associated with Bowker and Star's (1999) quote proposed earlier in the paper. Specifically, "everyday categories", which are reinforced by the functionality of generative technologies, arguably disappear and propagate through infrastructure, which is inaccessible due to its black-boxed nature, modularity and decentralisation (Bowker and Star, 1999, p.319).

Conclusion and Future Directions

In this position paper, I presented large-scale LMs and particularly GPT models as innovative, generative technologies, that disrupt traditional concepts of categorisation, infrastructure and infrastructure governance. At the core of my argument, GPTs are not solely artefacts, but non-neutral, complex systems that influence social reality and reinforce pre-existing orders and ideologies. The question I pose to the STS and SDS communities is, namely, how to update concepts of infrastructure governance accordingly, and rebalance the structural and infrastructural influence of generative technologies? While I will attempt to address this question through a proposed submission for the First Monday journal, here I presented some cues for the pursuit of this endeavour. Prominently, the theory of Social Worlds will be a starting point for this analysis, as it contemplates the co-existence and interaction of multiple social groups within and through sociotechnical systems (Clarke and Star, 2008). Additionally, in future efforts, I will suggest a potential, theoretical solution to the presented question based on the introduction of mixed qualitative-quantitative ontologies and epistemologies, with the objective of dismantling inflexible instances of infrastructure. Pragmatically, in the context of academic research, I will propose the recentring of research design practices around individuals and communities that are most affected by artefacts' functionalities. I will support such claims with a case study, as well as through the illustration of my current doctoral project.

References

1. Apple Inc., 2022. Siri does more than ever. Even before you ask.. (Accessed 23 February 2022).
2. Barocas, S., Hardt, M. & Narayanan, A., 2019. Fairness and Machine Learning: Limitations and Opportunities.
3. Basta, C., Costa-jussà, M. & Casas, N., 2019. *Evaluating the Underlying Gender Bias in Contextualized Word Embeddings*. Florence, Italy, Proceedings of the First Workshop on Gender Bias in Natural Language Processing.
4. Bowker, G., Baker, K., Millerand, F. & Ribes, D., 2010. Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In: *International Handbook of Internet Research*. London, United Kingdom: Springer, p. 97–117.
5. Bowker, G. & Star, S. L., 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, Massachusetts, United States : The MIT Press.
6. Carabantes, M., 2020. Black-Box Artificial Intelligence: an Epistemological and Critical Analysis. *AI & SOCIETY*, Volume 35, pp. 309-317.
7. Clarke, A. E. & Star, S. L., 2006. The Social Worlds Framework: A Theory/Methods Package. *The Handbook of Science and Technology Studies*, 3(0), pp. 113-137.
8. Costa, P. & Ribas, L., 2019. AI Becomes Her: Discussing Gender and Artificial Intelligence. *Technoetic Arts*, 17(1), pp. 171-193.
9. Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning*. Cambridge, Massachusetts: The MIT Press.
10. Harshvardhan, G. M., Gourisaria, M. K., Pandey, M. & Rautaray, S. S., 2020. A Comprehensive Survey and Analysis of Generative Models in Machine Learning. *Computer Science Review*, Volume 38, pp. 1-29.
11. Hester, H., 2017. Technology Becomes Her. *New Vistas*, 3(1), pp. 46-50.
12. Jurafsky, D. & Martin, J. H., 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 2rd ed. London, United Kingdom: Pearson.
13. Kurita, K. et al., 2019. *Measuring Bias in Contextualized Word Representations*. Florence, Italy, Proceedings of the First Workshop on Gender Bias in Natural Language Processing.
14. Li, H., 2022. Language Models: Past, Present, and Future. *Communications of the ACM*, 65(7), pp. 56-63.
15. Min, B. et al., 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. (Accessed 2 June 2022).
16. Nichols, J., Chan, H. H. & Baker, M., 2019. Machine Learning: Applications of Artificial Intelligence to Imaging and Diagnosis. *Biophysical Reviews*, 11(1), p. 111–118.
17. Noble, S., 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York City, New York, United States: New York University Press.
18. Offenhuber, D., 2017. *Waste is Information: Infrastructure Legibility and Governance*. Cambridge, Massachusetts, United States: The MIT Press.
19. Posner, M., 2018. See No Evil. *Logic*, 1 April.
20. Russell, S. & Norvig, P., 2010. *Artificial Intelligence: a Modern Approach*. 3rd ed. Harlow: Pearson Education Limited.
21. Sheng, E., Chang, K.-W., Natarajan, P. & Peng, N., 2019. *The Woman Worked as a Babysitter: On Biases in Language Generation*. Hong Kong, China, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.
22. Suzuki, M. & Matsuo, Y., 2022. A Survey of Multimodal Deep Generative Models. *Advanced Robotics*, 36(5-6), pp. 261-278.
23. Weiser, M., 1991. *Scientific American*, pp. 94-104.
24. Weizenbaum, J., 1976. *Computer Power and Human Reason: from Judgement to Calculations*. New York, San Francisco: W. H. Freeman and Company.