# COMPONENT 3

# An overview On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

## 1. BRIEF HISTORY OF LANGUAGE MODELS (LMS)

Language mode can be defined as a system trained to do string prediction. The statistical approach to string prediction was initially proposed by Dr Claude Shannon in 1949 but has been implemented for speech learning and machine translation in early 80's. In the past, we have seen a pattern of achieving better results through more data and increasing the size of models until scores don't see an improvement. New architectures are implemented that are able to take advantage of the big data we have currently. This has resulted in a change of the type of task these language models are used for.

## 2. RISKS OF LANGUAGE MODEL

### I. Costs

"The average human across the globe responsible for 5tons of CO2 emissions per year" (Strubell et al. 2019). They also looked at the process of training a transformer model and they found out that it would produce 284tons of CO2 emissions. To reduce carbon footprint, the authors gave recommendation to report the time taken to train sensitive hyperparameters and also urged government to invest in cloud computing to provide equitable access to researchers.

**Current mitigation efforts**

- Renewable energy sources: The world is opting for more cleaner sources of electricity but the disadvantage is that this still incurs a cost on the environment particularly in the form of infrastructure.

- Another mitigation strategy is to prioritize computationally efficient hardware as advocated by organizations such as Sustain NLP workshop. Schwartz et al. (2020) argue for promoting efficiency as the evaluation metric using green AI.

Large language models benefit hegemonic views but marginalised communities are most likely to be negatively impacted by climate change.

### II. Large Dataset

Several factors jeopardize internet participation. Think about the following:

Younger people and those from developed countries have more access to the web and thus are contributing more.

Twitter accounts receiving death threats more likely to be suspended due to twitter moderation than those issuing the threats.

What part of the internet are being included in these large datasets? Reddit-US users are 67 percent men and 64 percent are aged between 18 to 29. Wikipedia only 8.8 to 15 percent are women or girls. But not blog sites with fewer traffic.

Who are being filtered out? Some identities are being filtered primarily on target words referencing sex and also LGBTQ online spaces.

The reason this is a problem is that if we overrepresent a large amount of view point, then we are representing the language of people who are deliberately using their language in ways that are consistent with systems of oppression which can be expressed through racism, ageism, transphobia, etc.

Not only does the training data going to overrepresent hegemonic views, but the language model is going to absorb the biases from those views (Blodgett et al. 2020).

### III. Synthetic Data

**Stochastic**

From linguistics and psychology, we learn that human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996). In contrast, a language model is a system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning: *a stochastic parrot*.

The problem arises because as humans whenever we encounter synthetic text in a language that we are proficient in we make sense of it.

## 3. POTENTIAL HARMS

- Can be used for denigration, stereotype threat, hate speech: harms to reader, harms to bystanders.
- People can deliberately use these systems to create cheap synthetic text to do harm in the world like boosting extremist recruiting (McGuffie & Newhouse 2020).
- Language model errors attributed to human author in the other language.
- Language models can be probed to replicate personal identifiable information from the training data. (Carlini et al. 2020).
- Language models as hidden components can influence query expansion & results (Noble 2018).

## 4. CRITICAL REFLECTION

The paper answered questions surrounding if language models can be too big but does not give us an insight whether ever larger language models is inevitable or necessary?

Furthermore, the paper gives us more insight on the associated cost with this research direction but does not tell us what we should consider before pursuing it?

Lastly, it fails to inform us if NLP needs larger language models? Or how we can pursue mitigate the associated risks in this research direction?

Further research on the ethical and social risks of harm from language models (Weidinger, et al. 2021) throws more light on these questions but cannot be answered within the scope of this essay.

**References**


Bender, E. M., Gebru, T., McMillan-Major, A., et al. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In Proceedings of FAccT 2021.

Claude Elwood Shannon (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Emma Strubell, Ananya Ganesh, and Andrew McCallum (2019). *Energy and Policy Considerations for Deep Learning in NLP*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 3645–3650.

Herbert H. Clark (1996). *Using Language*. Cambridge University Press, Cambridge.

Kris McGuffie and Alex Newhouse (2020). *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*. Middlebury Institute of International Studies at Monterrey. https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf.

Nicholas Carlini, Florian Tramer, Eric Wallace, et al. (2020). *Extracting Training Data from Large Language Models*. arXiv:2012.07805 [cs.CR].

Roy Schwartz, Jesse Dodge, Noah A. Smith, et al. (2020). *Green AI*. ACM 63, 12 (Nov. 2020), 54–63. https://doi.org/10.1145/3381831

Safiya Umoja Noble (2018). *Algorithms of Oppression*: How Search Engines Reinforce Racism.

Seldon, (2021). *Explainability in machine learning.* Available at: https://www.seldon.io/explainability-in-machine-learning (Accessed at: 10th May 2022).

Steven J. Bowen (2017). *Total value optimization.* SJDB LLC.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, et al. (2020). *Language (Technology) is Power: A Critical Survey of bias in NLP*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 5454–5476. https: //doi.org/10.18653/v1/2020.acl-main.485.

Weidinger, L., Mellor, J., Rauh, M., et al. (2021). *Ethical and social risks of harm from language models*. https://arxiv.org/abs/2112.04359.