**Bias and Toxicity**

Bias, as defined by [Merriam-Webster](https://www.merriam-webster.com/dictionary/bias), is "an inclination of temperament or outlook…especially: a personal and sometimes unreasoned judgement." This definition of bias, particularly the words "unreasoned judgement," suggest that someone's biased outlook could be harmful towards someone. Now, let's compare bias to toxicity. Toxicity, on the other hand, is much more forthcoming about its harmfulness. [Merriam-Webster](https://www.merriam-webster.com/dictionary/toxicity) defines it as "the quality, state, or relative degree of being poisonous." While bias will always be present in human thoughts and opinions, toxicity has the capacity of turning bias into something much more powerful.

Large Language Models (LLMs) have both bias and toxicity from their immense amounts of training data. It's no surprise that LLMs would pick up our human "unreasoned judgements" and "poisonous" thoughts all over the Internet. Canonical [research](https://dl.acm.org/doi/10.1145/3442188.3445922) on LLMs has found that even the Internet itself is biased. Due to issues like Internet access, it mainly captures data from younger users and people from developed countries. The result? LLMs are *not* accounting - at least not sufficiently - for the data of those being marginalized. In other words, LLMs perpetuate dominant and stereotypical narratives about society, people, and culture. In our country, this means worsening political and socio-economic divides, including the flaring of especially toxic points of view related to sex, race, and immigration. Ultimately, It doesn't take much to go online to one of these models and think it's confirming what you already believe.

To make matters worse, the present bias and toxicity in LLMs only increases with the size of the model. Recently, [researchers](https://www.technologyreview.com/2023/03/20/1070067/language-models-may-be-able-to-self-correct-biases-if-you-ask-them-to/) have learned that we can ask certain models to "self-correct" on their bias, which has yielded successful results. The optimal solution would be to train models to not be biased or toxic, but even researchers working on the "self-correction" point to this as being a larger, systemic issue - and, I agree. Other professionals, beyond engineers and artificial intelligence (AI) firms, need to be involved in making LLMs safer and more inclusive. Open AI [put out a call](https://openai.com/research/ai-safety-needs-social-scientists) back in 2019 for social scientists to help align AI to human values. But *what* are our human values, especially in our "us versus them" climate? Do we need to leverage the help of the government for some regulations, or would that just take too long? Maybe bringing it back to a more grassroots level could help too, even if it simply starts with raising more awareness among those being impacted the most. Populations being marginalized by LLMs may be too busy trying to survive, providing for themselves as well as for their families, to know and understand the full extent and repercussions of the bias and toxicity in LLMs - now and in the future. Maybe the time to act is now more than ever.