Implicit Bias in Language Models – A Narrative Literature Review with Systematic Elements

Kevin Kiy
Maynooth University, Maynooth, Ireland

Dermot Lynott
Maynooth University, Maynooth, Ireland

Diarmuid O'Donoghue
Maynooth University, Maynooth, Ireland

Abstract

Implicit biases are a common source of prejudicial decision-making in society. While the use of language models might be intended to eliminate human bias and prevent harmful prejudice, they are trained on human-generated linguistic data and thus inherit human-like biased attitudes. We conducted a narrative review of implicit attitudes in linguistic models, drawing on literature from artificial intelligence, social psychology, and cognitive science. Drawn from experimental data, our findings suggest an important link between statistical patterns in language and the implicit biases displayed by people. While several efforts have been made to capture the levels of bias in language models, there is no contribution yet that focuses on the causal nature of the relationship between language and implicit bias in language models. This literature review highlights the state of the art in this growing field, identifies gaps in the literature, and showcases challenges for further research in the future.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*