HOMOPHOBIA OF AI

AI LANGUAGE MODELS

LANGUAGE MODELS ANALYZE BODIES OF TEXT DATA TO PROVIDE WORD PREDICTIONS IN SENTENCES, USING THE INTERNET TO TRAIN THEM. BECAUSE THIS IS DERIVED FROM THE INTERNET, AIS ARE STARTING TO ADOPT HATE SPEECH OF ALL KINDS SUCH AS ANTI-QUEER BIAS.

FELKNER

Katy Felkner, a Ph.D. student in computer science, developed a benchmark for measuring bias in AI called Felkner. In project research, a popular model called BERT showed high levels of homophobic bias according to the Felkner benchmark. It predicts the usage of heteronormativity 74% of the time instead of queer inclusion.

BIAS BENCHMARKS

The Felkner benchmark let researchers know to retrain BERT and fed the model more queer content bringing the bias score from 74% to 55%. Bias benchmarks will help change large AI languages by bringing awareness to these researchers to train these models for the better not only for homophobia bias but for racism, ableism, and more.