As AI tools get smarter, they’re growing more covertly racist, experts find

Publication Date: 2024-03-16

Author: Ava Sasani

Section: Technology

Tags: Artificial intelligence (AI), Race, OpenAI, ChatGPT, Google, Alphabet, Computing, news

Article URL: https://www.theguardian.com/technology/2024/mar/16/ai-racism-chatgpt-gemini-bias

Popular artificial intelligence tools are becoming more covertly racist as they advance, says an alarming new report. A team of technology and linguistics researchers revealed this week that large language models like OpenAI’s ChatGPT and Google’s Gemini hold racist stereotypes about speakers of African American Vernacular English, or AAVE, an English dialect created and spoken by Black Americans. “We know that these technologies are really commonly used by companies to do tasks like screening job applicants,” said Valentin Hoffman, a researcher at the Allen Institute for Artificial Intelligence and co-author of the recent paper, published this week in arXiv, an open-access research archive from Cornell University. Hoffman explained that previously researchers “only really looked at what overt racial biases these technologies might hold” and never “examined how these AI systems react to less overt markers of race, like dialect differences”. Black people who use AAVE in speech, the paper says, “are known to experience racial discrimination in a wide range of contexts, including education, employment, housing, and legal outcomes”. Hoffman and his colleagues asked the AI models to assess the intelligence and employability of people who speak using AAVE compared to people who speak using what they dub “standard American English”. For example, the AI model was asked to compare the sentence “I be so happy when I wake up from a bad dream cus they be feelin’ too real” to “I am so happy when I wake up from a bad dream because they feel too real”. The models were significantly more likely to describe AAVE speakers as “stupid” and “lazy”, assigning them to lower-paying jobs. Hoffman worries that the results mean that AI models will punish job candidates for code-switching – the act of altering how you express yourself based on your audience – between AAVE and standard American English. “One big concern is that, say a job candidate used this dialect in their social media posts,” he told the Guardian. “It’s not unreasonable to think that the language model will not select the candidate because they used the dialect in their online presence.” The AI models were also significantly more likely to recommend the death penalty for hypothetical criminal defendants that used AAVE in their court statements. “I’d like to think that we are not anywhere close to a time when this kind of technology is used to make decisions about criminal convictions,” said Hoffman. “That might feel like a very dystopian future, and hopefully it is.” Still, Hoffman told the Guardian, it is difficult to predict how language learning models will be used in the future. “Ten years ago, even five years ago, we had no idea all the different contexts that AI would be used today,” he said, urging developers to heed the new paper’s warnings on racism in large language models. Notably, AI models are already used in the US legal system to assist in administrative tasks like creating court transcripts and conducting legal research. For years, leading AI experts like Timnit Gebru, former co-leader of Google’s ethical artificial intelligence team, have called for the federal government to curtail the mostly unregulated use of large language models. “It feels like a gold rush,” Gebru told the Guardian last year. “In fact, it is a gold rush. And a lot of the people who are making money are not the people actually in the midst of it.” Google’s AI model, Gemini, found itself in hot water recently when a slew of social media posts showed its image generation tool depicting a variety of historical figures – including popes, founding fathers of the US and, most excruciatingly, German second world war soldiers – as people of color. Large language models improve as they are fed more data, learning to more closely mimic human speech by studying text from billions of web pages across the internet. The long-acknowledged conceit of this learning process is that the model will spew whatever racist, sexist, and otherwise harmful stereotypes it encounters on the internet: in computing, this problem is described by the adage “garbage in, garbage out”. Racist input leads to racist output, causing early AI chatbots like Microsoft’s Tay to regurgitate the same neo-Nazi content it learned from Twitter users in 2016. In response, groups like OpenAI developed guardrails, a set of ethical guidelines that regulate the content that language models like ChatGPT can communicate to users. As language models become larger, they also tend to become less overtly racist. But Hoffman and his colleagues found that, as language models grow, covert racism increases. Ethical guardrails, they learned, simply teach language models to be more discreet about their racial biases. “It doesn’t eliminate the underlying problem; the guardrails seem to emulate what educated people in the United States do,” said Avijit Ghosh, an AI ethics researcher at Hugging Face, whose work focuses on the intersection of public policy and technology. “Once people cross a certain educational threshold, they won’t call you a slur to your face, but the racism is still there. It’s a similar thing in language models: garbage in, garbage out. These models don’t unlearn problematic things, they just get better at hiding it.” The US private sector’s open-armed embrace of language models is expected to intensify over the next decade: the broader market of generative AI is projected to become a $1.3tn industry by 2032, according to Bloomberg. Meanwhile, federal labor regulators like the Equal Employment Opportunity Commission only recently began shielding workers from AI-based discrimination, with the first case of its kind coming before the EEOC late last year. Ghosh is part of the growing contingent of AI experts who, like Gebru, worry about the harm the language learning models might cause if technological advancements continue to outpace federal regulation. “You don’t need to stop innovation or slow AI research, but curtailing the use of these technologies in certain sensitive areas is an excellent first step,” he said. “Racist people exist all over the country; we don’t need to put them in jail, but we try to not allow them to be in charge of hiring and recruiting. Technology should be regulated in a similar way.”