Programs to detect AI discriminate against non-native English speakers, shows study

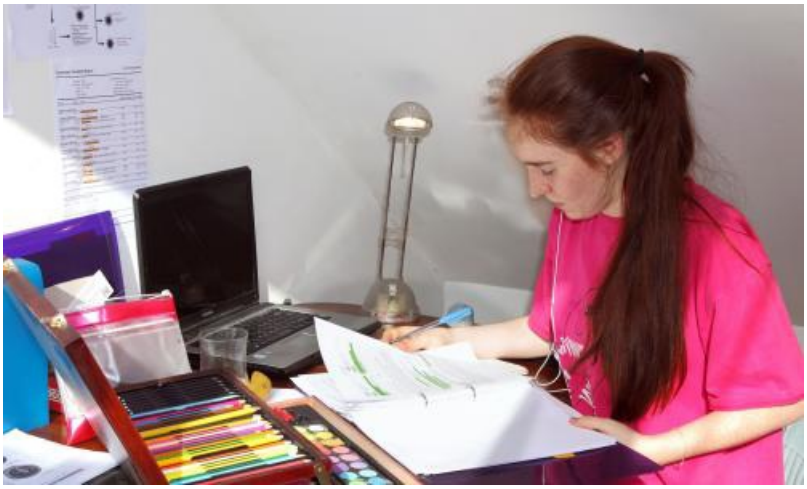
Publication Date: 2023-07-10

Author: Ian Sample

Section: Technology

Tags: Artificial intelligence (AI), ChatGPT, Exams, Computing, news

Article URL: https://www.theguardian.com/technology/2023/jul/10/programs-to-detect-ai-discriminate-against-non-native-english-speakers-shows-study



Computer programs that are used to detect essays, job applications and other work generated by artificial intelligence can discriminate against people who are non-native English speakers, researchers say. Tests on seven popular AI text detectors found that articles written by people who did not speak English as a first language were often wrongly flagged as AI-generated, a bias that could have a serious impact on students, academics and job applicants. With the rise of ChatGPT, a generative AI program that can write essays, solve problems and create computer code, many teachers now consider AI detection as a “critical countermeasure to deter a 21st-century form of cheating”, the researchers say, but they warn that the 99% accuracy claimed by some detectors is “misleading at best.” Scientists led by James Zou, an assistant professor of biomedical data science at Stanford University, ran 91 English essays written by non-native English speakers through seven popular GPT detectors to see how well the programs performed. More than half of the essays, which were written for a widely recognised English proficiency test known as the Test of English as a Foreign Language, or TOEFL, were flagged as AI-generated, with one program flagging 98% of the essays as composed by AI. When essays written by native English-speaking eighth graders in the US were run through the programs, the same AI detectors classed more than 90% as human-generated. Writing in the journal Patterns, the scientists traced the discrimination to the way the detectors assess what is human and what is AI-generated. The programs look at what is called “text perplexity”, which is a measure of how “surprised” or “confused” a generative language model is when trying to predict the next word in a sentence. If the model can predict the next word easily, the text perplexity is ranked low, but if the next word proves hard to predict, the text perplexity is rated high. Large language models or LLMs like ChatGPT are trained to churn out low perplexity text, but this means that if humans use a lot of common words in a familiar pattern in their writing, their work is at risk of being mistaken for AI-generated text. The risk is greater with non-native English speakers, the researchers say, because they are more likely to adopt simpler word choices. After highlighting the built-in bias in the AI detector programs, the scientists went back to ChatGPT and asked it to rewrite the TOEFL essays using more sophisticated language. When these edited essays were run back through the AI detectors, they were all labelled as written by humans. “Paradoxically, GPT detectors might compel non-native writers to use GPT more to evade detection,” they said. “The implications of GPT detectors for non-native writers are serious, and we need to think through them to avoid situations of discrimination,” the authors warned in the journal. AI detectors could falsely flag college and job applications as GPT-generated, and marginalise non-native English speakers on the internet, because search engines such as Google downgrade what is assessed to be AI-generated content, they warn. “In education, arguably the

most significant market for GPT detectors, non-native students bear more risks of false accusations of cheating, which can be detrimental to a student's academic career and psychological wellbeing," the researchers added. In an accompanying article, Jahna Otterbacher at the Cyprus Center for Algorithmic Transparency at the Open University of Cyprus, said: "Rather than fighting AI with more AI, we must develop an academic culture that promotes the use of generative AI in a creative, ethical manner … ChatGPT is constantly collecting data from the public and learning to please its users; eventually, it will learn to outsmart any detector."