# As AI cheating booms, so does the industry detecting it: ‘We couldn’t keep up with demand’

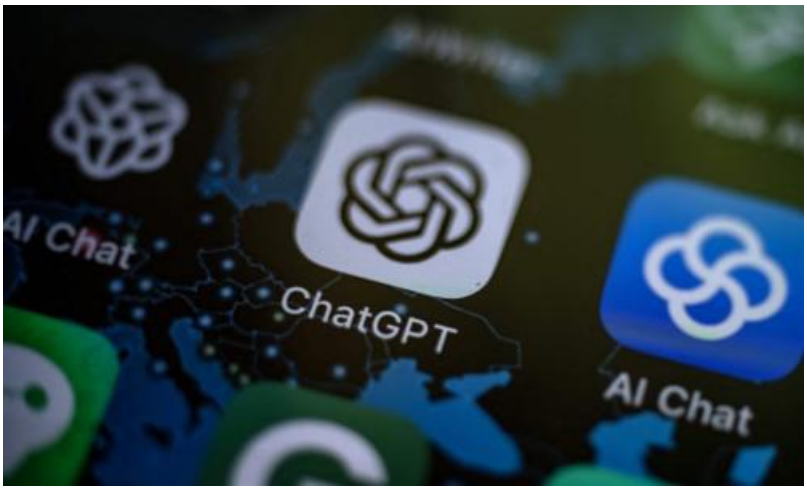
Publication Date: 2023-07-05

Author: Unknown

Section: Technology

Tags: Artificial intelligence (AI), Computing, features

Article URL: https://www.theguardian.com/technology/2023/jul/05/as-ai-cheating-booms-so-does-the-industry-detecting-it-we-couldnt-keep-up-with-demand




Since its release last November, ChatGPT has shaken the education world. The chatbot and other sophisticated AI tools are reportedly being used everywhere from college essays to high school art projects. A recent survey of 1,000 students at four-year universities by Intelligent.com found that 30% of college students have reported using ChatGPT on written assignments. This is a problem for schools, educators and students – but a boon for a small but growing cohort of companies in the AI-detection business. Players like Winston AI, Content at Scale and Turnitin are billing for their ability to detect AI-involvement in student work, offering subscription services where teachers can run their students’ work through a web dashboard and receive a probability score that grades how “human” or “AI” the text is. At this stage, most clients are teachers acting on their own initiative, although Winston AI says it is beginning talks with school administrators at the district level as the problem grows. And with only one full academic semester since ChatGPT was released, the disruption and headaches are only beginning. Methods for detecting AI-generated content typically involve the search for a “tell” – a feature that distinguishes an AI author from a human one. According to MIT Technology Review’s guide, in AI content “the word ‘the’ can occur too many times”. The text can also have a sort of anti-style indicating a lack of human flair. The presence of typos is often a dead giveaway for a human mind – LLMs (large language models like ChatGPT) have Scripps Spelling Bee-winning skills. Visual generative AI has its own teething issues; mistakes like a hand with too many fingers are common. AI relies on patterns and phrases in its training data – just like the problem of overusing the word “the”, sometimes it can rely on these patterns too much. John Renaud, the co-founder of Winston AI, says two of the most notable tells they’re looking for are “perplexity” and “burstiness”. “Perplexity” refers to the sophistication of language patterns that appear within a text sample (is this a pattern that exists in the training data, or is it intricate enough to seem novel?), while “burstiness” refers to “when a text features a cluster of words and phrases that are repeated within a short span of time”. Renaud says the company saw a surge of interest in the wake of ChatGPT: “It all happened within a week or two – suddenly we couldn’t keep up with demand.” And it’s not just academia: school essays are the most commonly scanned content but the second “would be publishers scanning their journalists’/copywriters’ work before publishing”. The company claims to be one of the more accurate detectors around, boasting a 99.6% accuracy rate. Even though he was “very worried” about ChatGPTs initial breakout, Renaud has since become more sanguine. “With predictive AI, we’ll always be able to build a model to predict it,” he says. In other words, the current generation of autocomplete-on-steroids algorithms will always be deterministic enough to have tells. Annie Chechitelli, Turnitin’s chief product officer, also thinks AI fears are overblown, publishing a letter recently to the Chronicle

of Higher Education titled “Not True That ChatGPT Can’t Be Accurately Detected” and pushing back on claims we’ve gone through the generated-content looking-glass. “We think there will always be a tell,” she says over Zoom. “And we’re seeing other methods to unmask it. We have cases now where teachers want students to do something in person to establish a baseline. And keep in mind that we have 25 years of student data to train our model on.” And like Renaud at Winston AI, Chechitelli is seeing an explosion of interest in her services and AI detection in general. “A survey is conducted every year of teachers’ top instructional challenges. In 2022 ‘preventing student cheating’ was 10th,” she says. “Now it’s number one.” Altogether, the state of the industry gives the impression of an arms race between AI generators and AI detectors lasting years, each trading supremacy as the technological tit-for-tat plays out. While some believe humans will remain one step ahead, others are more bearish about the potential for these tools to eventually avoid our detection. Irene Solaiman, policy director at AI startup Hugging Face, recently wrote in the MIT Technology Review: “The bigger and more powerful the model, the harder it is to build AI models to detect what text is written by a human and what isn’t.” One larger solution that’s being proposed is “watermarks”. The idea is that models such as ChatGPT could be made to structure sentences in ways that identify that the content is AI generated, deliberately inserting the “tells” that detection software is already looking for. But both Chechitelli and Renaud agree that the idea has flaws, especially if it is not universally adopted. If there were an alternative, “everyone is just gonna flock to the one without the watermark,” Renaud says. Why would someone use an algorithm that tattled on them, versus one that just quietly produced convincing content? The era of the human-authored web is ending, and no one is entirely sure what comes next. Whether AI content becomes indistinguishable or the human touch proves impossible to replicate, one thing is certain – there will be power for those who can tell the difference.