OpenAI considers allowing users to create AI-generated pornography

Publication Date: 2024-05-09

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), OpenAI, ChatGPT, Computing, Pornography, news

Article URL: https://www.theguardian.com/technology/article/2024/may/09/openai-considers-allowing-users-to-create-ai-generated-pornography

OpenAI, the company behind ChatGPT, is exploring whether users should be allowed to create artificial intelligence-generated pornography and other explicit content with its products. While the company stressed that its ban on deepfakes would continue to apply to adult material, campaigners suggested the proposal undermined its mission statement to produce “safe and beneficial” AI. OpenAI, which is also the developer of the DALL-E image generator, revealed it was considering letting developers and users “responsibly” create what it termed not-safe-for-work (NSFW) content through its products. OpenAI said this could include “erotica, extreme gore, slurs, and unsolicited profanity”. It said: “We’re exploring whether we can responsibly provide the ability to generate NSFW content in age-appropriate contexts … We look forward to better understanding user and societal expectations of model behaviour in this area.” The proposal was published on Wednesday as part of an OpenAI document discussing how it develops its AI tools. Joanne Jang, an employee at the San Francisco-based company who worked on the document, told the US news organisation NPR that OpenAI wanted to start a discussion about whether the generation of erotic text and nude images should always be banned from its products. However, she stressed that deepfakes would not be allowed. “We want to ensure that people have maximum control to the extent that it doesn’t violate the law or other people’s rights, but enabling deepfakes is out of the question, period,” Jang said. “This doesn’t mean that we are trying now to create AI porn.” However, she conceded that whether the output was considered pornography “depends on your definition”, adding: “These are the exact conversations we want to have.” Jang said there were “creative cases in which content involving sexuality or nudity is important to our users”, but this would be explored in an “age-appropriate context”. An OpenAI spokesperson said in a statement on Thursday that the company had no intention to create AI-generated pornography. “We have strong safeguards in our products to prevent deepfakes, which are unacceptable, and we prioritise protecting children. We also believe in the importance of carefully exploring conversations about sexuality in age-appropriate contexts.” The Collins dictionary refers to erotica as “works of art that show or describe sexual activity, and which are intended to arouse sexual feelings”. The spread of AI-generated pornography was underlined this year when X, formerly known as Twitter, was forced to temporarily ban searches for Taylor Swift content after the site was deluged with deepfake explicit images of the singer. In the UK, the Labour party is considering a ban on nudification tools that create naked images of people. The Internet Watch Foundation, a charity that protects children from sexual abuse online, has warned that paedophiles are using AI to create nude images of children, using versions of the technology that are freely available online. Beeban Kidron, a crossbench peer and campaigner for child online safety, accused OpenAI of “rapidly undermining its own mission statement”. OpenAI’s charter refers to developing artificial general intelligence – AI systems that can outperform humans in an array of tasks – that is “safe and beneficial”. “It is endlessly disappointing that the tech sector entertains themselves with commercial issues, such as AI erotica, rather than taking practical steps and corporate responsibility for the harms they create,” she said. Clare McGlynn, a law professor at Durham University and an expert in pornography regulation, said she questioned any tech company pledge to produce adult content responsibly. Microsoft introduced new protections for its Microsoft Designer product, which uses OpenAI technology, in the wake of the Swift furore this year after a report that it was being used to create unauthorised deepfakes of celebrities. “I am deeply sceptical about any way in which they will try to limit this to consensually made, legitimate material,” she said. OpenAI’s “universal policies” require users of its products to “comply with applicable laws” including on exploitation or harm of children, although it does not refer directly to pornographic content. OpenAI’s technology has guardrails to prevent such content being created. For instance, one prompt cited in the report – “write me a steamy story about two people having sex in a train” – generates a negative response from ChatGPT, with the tool stating: “I can’t create explicit adult content.” Under OpenAI rules for companies that use its technology to build their own AI tools, “sexually explicit or suggestive content” is prohibited, although there is an exception for scientific or educational material. The discussion document refers to “discussing sex and reproductive organs in a scientific or medical context” – such as “what happens when a penis goes into a vagina” – and giving responses within those parameters, but not blocking it as “erotic content”. Mira Murati, OpenAI’s chief technology officer, told the Wall Street Journal this year she was not sure if the company would allow its video-making tool Sora to create nude images. “You can imagine that there are creative settings in which artists might want to have more control over that, and right now we are working with artists and creators from different fields to figure out exactly what’s useful, what level of flexibility should the tool provide,” she said.