Google pauses AI-generated images of people after ethnicity criticism

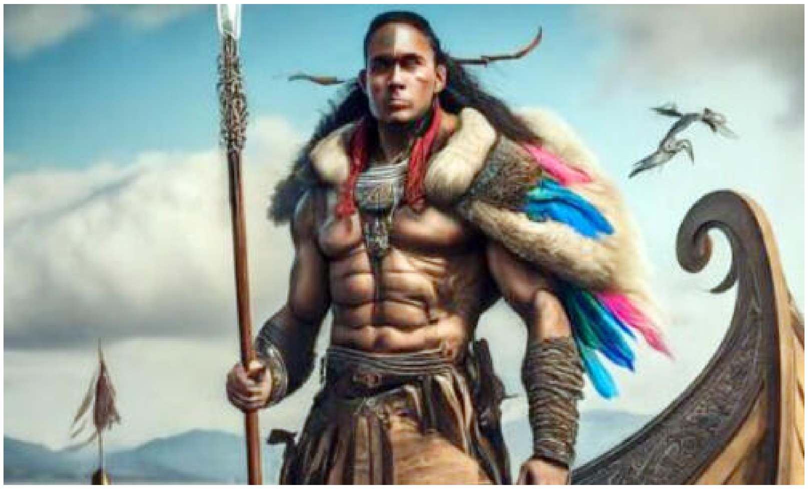
Publication Date: 2024-02-22

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Google, Race, news

Article URL: https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism



Google has put a temporary block on its new artificial intelligence model producing images of people after it portrayed German second world war soldiers and Vikings as people of colour. The tech company said it would stop its Gemini model generating images of people after social media users posted examples of images generated by the tool that depicted some historical figures – including popes and the founding fathers of the US – in a variety of ethnicities and genders. “We’re already working to address recent issues with Gemini’s image generation feature. While we do this, we’re going to pause the image generation of people and will rerelease an improved version soon,” Google said in a statement. Google did not refer to specific images in its statement, but examples of Gemini image results were widely available on X, accompanied by commentary on AI’s issues with accuracy and bias, with one former Google employee saying it was “hard to get Google Gemini to acknowledge that white people exist”. Jack Krawczyk, a senior director on Google’s Gemini team, had admitted on Wednesday that the model’s image generator – which is not available in the UK and Europe – needed adjustment. “We’re working to improve these kinds of depictions immediately,” he said. “Gemini’s AI image generation does generate a wide range of people. And that’s generally a good thing because people around the world use it. But it’s missing the mark here.” Krawczyk added in a statement on X that Google’s AI principles committed its image generation tools to “reflect our global user base”. He added that Google would continue to do this for “open ended” image requests such as “a person walking a dog” but acknowledged that the response prompts with a historical slant needed further work. “Historical contexts have more nuance to them and we will further tune to accommodate that,” he said. Coverage of bias in AI has shown numerous examples of a negative impact on people of colour. A Washington Post investigation last year showed multiple examples of image generators showing bias against people of colour, as well as sexism. It found that the image generator Stable Diffusion XL showed recipients of food stamps as being primarily non-white or darker-skinned despite 63% of the recipients of food stamps in the US being white. A request for an image of a person “at social services” produced similar results. Andrew Rogoyski, of the Institute for People-Centred AI at the University of Surrey, said it was a “hard problem in most fields of deep learning and generative AI to mitigate bias” and mistakes were likely to occur as a result. “There is a lot of research and a lot of different approaches to eliminating bias, from curating training datasets to introducing guardrails for trained models,” he said. “It’s likely that AIs and LLMs [large language models] will continue to make mistakes but it’s also likely that this will improve over time.”