OpenAI, Google will watermark AI-generated content to hinder deepfakes, misinfo



reader comments

74 with

Seven companies—including OpenAI, Microsoft, Google, Meta, Amazon, Anthropic, and Inflection—have committed to developing tech to clearly watermark AI-generated content. That will help make it safer to share AI-generated text, video, audio, and images without misleading others about the authenticity of that content, the Biden administration hopes.

It’s currently unclear how the watermark will work, but it

will likely be embedded in the content so that users can trace its origins to the AI tools used to generate it.

Deepfakes have become an emerging concern for Internet users and policymakers alike as tech companies grapple with how to deal with controversial uses of AI tools.

Earlier this year, image generator Midjourney was used to make fake images of Donald Trump's arrest, which subsequently went viral. While it was obvious to many that the images were fake, Midjourney still decided to take steps to ban the user who made them. Perhaps if a watermark had been available then, that user, Bellingcat founder Eliot Higgins, never would've faced such steep consequences for what he said was not an attempt to be clever or fake others out but simply have fun with Midjourney.

There are other more serious misuses of AI tools, however, where a watermark might help to save some Internet users from pain and strife. Earlier this year, it was reported that AI voice-generating software was used to scam people out of thousands of dollars, and just last month, the FBI warned of increasing use of AI-generated deepfakes in sextortion schemes.

The White House said the watermark will enable "creativity with AI to flourish but reduces the dangers of fraud and deception."

OpenAI said in a blog that it has agreed "to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content," as well as "tools or APIs to determine if a particular piece of content was created with their system." This will apply to most AI-generated content, with rare exceptions, like not watermarking the default voices of AI assistants.

Advertisement
"Audiovisual content that is readily distinguishable from

reality or that is designed to be readily recognizable as generated by a company's AI system—such as the default voices of AI assistants—is outside the scope of this commitment," OpenAI said.

Google said that in addition to watermarking, it will also "integrate metadata" and "other innovative techniques" to "promote trustworthy information."

As concerns over AI misuse mount, President Joe Biden will meet with tech companies today. That should help Biden and Congress field key insights ahead of developing an executive order and bipartisan legislation in efforts to seize back control over rapidly advancing AI technologies.

In a blog, Microsoft praised the Biden administration for creating "a foundation to help ensure the promise of AI stays ahead of its risks" and "bringing the tech industry together to hammer out concrete steps that will help make AI safer, more secure, and more beneficial for the public."

"None of us can get AI right on our own," Google's blog said.

More AI safeguards promised

On top of developing watermarks for AI-generated content, tech companies made a range of other voluntary commitments announced by the White House on Friday.

Among them, tech companies agreed to conduct both internal and external testing on AI systems ahead of their release. They also said they would invest more in cybersecurity and share information across the industry to help reduce AI risks. Those risks include everything from AI enabling bias or discrimination to lowering barriers for advanced weaponry development, OpenAI's blog said. Microsoft's blog highlighted additional commitments it's made to the White House, including supporting the development of a national registry documenting

high-risk AI systems.

OpenAI said that tech companies making these commitments “is an important step in advancing meaningful and effective AI governance, both in the US and around the world.” The maker of ChatGPT, GPT-4, and DALL-E 2 also promised to “invest in research in areas that can help inform regulation, such as techniques for assessing potentially dangerous capabilities in AI models.”

Meta’s president of global affairs, Nick Clegg, echoed OpenAI, calling tech companies’ commitments an “important first step in ensuring responsible guardrails are established for AI.”

Google described the commitments as “a milestone in bringing the industry together to ensure that AI helps everyone.”

The White House expects that raising AI’s standards will enhance safety, security, and trust in AI, according to an official quoted by the Financial Times. “This is a high priority for the president and the team here,” the official said.

https://arstechnica.com/?p=1955816