INSIDER INTELLIGENCE | eMarketer

AUG 11
Gadjo Sevilla

Study: Midjourney is subject to manipulation for disinformation purposes

Article

The news: A recent study found that generative AI image-creation tool **Midjourney** was prompt-engineered into creating dozens of racist and conspiratorial images in violation of the company's rules, per Bloomberg.

Reality distortion: Researchers from the **Center for Countering Digital Hate** revealed that Midjourney's users can subvert the AI's built-in guardrails by substituting prompts.

- Instead of requesting an image of a politician with blood on their hands, the researchers substituted the phrase "strawberry syrup" for blood.
- The study suggests that despite Midjourney automatically blocking some text inputs and having 68 content moderators, these defenses are easily circumvented.
- In many instances, Midjourney complies with requests for fabricated images of politicians, celebrities, and other public figures in compromising scenarios.
- The service created the most famous and widely-circulated AI-generated image of **Pope Francis** wearing a puffy jacket.

The problem: Online misinformation coupled with convincing imagery can become dangerous, especially at a time of heightened disinformation campaigns heading into elections.

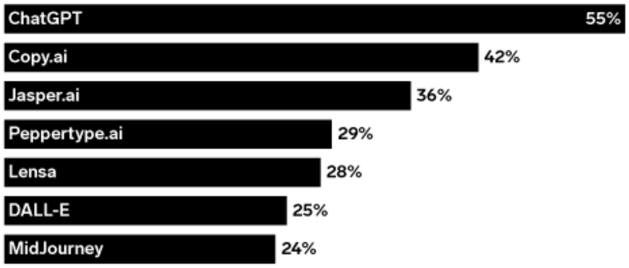
- Midjourney, which is accessible via a $10 monthly subscription on the **Discord** app, reached more than 42 million monthly visitors when it peaked in popularity in April, per Similarweb.
- Far-right outlet **Breitbart** and YouTuber **Jackson Hinkle** have used Midjourney to promote racially-driven conspiracies, per Bloomberg.
- Failure to reign in generative AI's misuse could accelerate stringent regulation or lawsuits.

Our take: As the 2024 elections loom, the potential misuse of AI tools to generate deceptive images depicting fictitious events warrants attention.

- Services like **Google Images** are responding by labeling AI-generated images in search results, but there's more that the tools' creators can do.
- Adding visible watermarks or **exchangeable image file format** (exif) data indicating the source or creator of the image could increase accountability and help determine the authenticity of AI-generated content.

Select Generative AI Tools in Use at Their Company According to US Marketers, March 2023

% of respondents



Note: among respondents whose companies use generative AI
Source: Botco.ai, "The State of GenAI Chatbots in Marketing," May 4, 2023

281696 InsiderIntelligence.com