FROM SHRIMP JESUS TO FAKE SELF-PORTRAITS, AI-GENERATED IMAGES HAVE BECOME THE LATEST FORM OF SOCIAL MEDIA SPAM - THE CONVERSATION

Posted on April 29, 2024 by Renee DiResta, Abhiram Reddy, and Josh A. Goldstein

Many of the AI images generated by spammers and scammers have religious themes.
immortal70/iStock via Getty Images

Renee DiResta, *Stanford University*; Abhiram Reddy, *Georgetown University*, and Josh A. Goldstein, *Georgetown University*

If you've spent time on Facebook over the past six months, you may have noticed photorealistic images that are too good to be true: children holding paintings that look like the work of professional artists, or majestic log cabin interiors that are the stuff of Airbnb dreams.

Others, such as renderings of Jesus made out of crustaceans, are just bizarre.

Like the AI image of the pope in a puffer jacket that went viral in May 2023, these AI-generated images are increasingly prevalent – and popular – on social media platforms. Even as many of them border on the surreal, they're often used to bait engagement from ordinary users.

Our team of researchers from the Stanford Internet Observatory and Georgetown University's Center for Security and Emerging Technology investigated over 100 Facebook pages that posted high volumes of AI-generated content. We published the results in March 2024 as a preprint paper, meaning the findings have not yet gone through peer review.

We explored patterns of images, unearthed evidence of coordination between some of the pages, and tried to discern the likely goals of the posters.

Page operators seemed to be posting pictures of AI-generated babies, kitchens or birthday cakes for a range of reasons.

There were content creators innocuously looking to grow their followings with synthetic content; scammers using pages stolen from small businesses to advertise products that don't seem to exist; and spammers sharing AI-generated images of animals while referring users to websites filled with advertisements, which allow the owners to collect ad revenue without creating high-quality content.

Our findings suggest that these AI-generated images draw in users – and Facebook's recommendation algorithm may be organically promoting these posts.

Generative AI meets scams and spam

Internet spammers and scammers are nothing new.

For more than two decades, they've used unsolicited bulk email to promote pyramid schemes. They've targeted senior citizens while posing as Medicare representatives or computer technicians.

On social media, profiteers have used clickbait articles to drive users to ad-laden websites. Recall the 2016 U.S. presidential election, when Macedonian teenagers shared sensational political memes on Facebook and collected advertising revenue after users visited the URLs they posted. The teens didn't care who won the election. They just wanted to make a buck.

In the early 2010s, spammers captured people's attention with ads promising that anyone could lose belly fat or learn a new language with "one weird trick."

AI-generated content has become another "weird trick."

It's visually appealing and cheap to produce, allowing scammers and spammers to generate high volumes of engaging posts. Some of the pages we observed uploaded dozens of unique images per day. In doing so, they followed Meta's own advice for page creators. Frequent posting, the company suggests, helps creators get the kind of algorithmic pickup that leads their content to appear in the "Feed," formerly known as the "News Feed."

Much of the content is still, in a sense, clickbait: Shrimp Jesus makes people pause to gawk and inspires shares purely because it is so bizarre.

Many users react by liking the post or leaving a comment. This signals to the algorithmic curators that perhaps the content should be pushed into the feeds of even more people.

Some of the more established spammers we observed, likely recognizing this, improved their engagement by pivoting from posting URLs to posting AI-generated images. They would then comment on the post of the AI-generated images with the URLs of the ad-laden content farms they wanted users to click.

But more ordinary creators capitalized on the engagement of AI-generated images, too, without obviously violating platform policies.

Rate 'my' work!

When we looked up the posts' captions on CrowdTangle – a social media monitoring platform owned by Meta and set to sunset in August – we found that they were "copypasta" captions, which means that they were repeated across posts.

Some of the copypasta captions baited interaction by directly asking users to, for instance, rate a "painting" by a first-time artist – even when the image was generated by AI – or to wish an elderly person a happy birthday. Facebook users often replied to AI-generated images with comments of encouragement and congratulations

Algorithms push AI-generated content

Our investigation noticeably altered our own Facebook feeds: Within days of visiting the pages – and without commenting on, liking or following any of the material – Facebook's algorithm recommended reams of other AI-generated content.

Interestingly, the fact that we had viewed clusters of, for example, AI-generated miniature cow pages didn't lead to a short-term increase in recommendations for pages focused on actual miniature cows, normal-sized cows or other farm animals. Rather, the algorithm recommended pages on a range of topics and themes, but with one thing in common: They contained AI-generated images.

In 2022, the technology website Verge detailed an internal Facebook memo about proposed changes to the company's algorithm.

The algorithm, according to the memo, would become a "discovery-engine," allowing users to come into contact with posts from individuals and pages they didn't explicitly seek out, akin to TikTok's "For You" page.

We analyzed Facebook's own "Widely Viewed Content Reports," which lists the most popular content, domains, links, pages and posts on the platform per quarter.

It showed that the proportion of content that users saw from pages and people they don't follow steadily increased between 2021 and 2023. Changes to the algorithm have allowed more room for AI-generated content to be organically recommended without prior engagement – perhaps explaining our experiences and those of other users.

'This post was brought to you by AI'

Since Meta currently does not flag AI-generated content by default, we sometimes observed users warning others about scams or spam AI content with infographics.

Meta, however, seems to be aware of potential issues if AI-generated content blends into the information environment without notice. The company has released several announcements about how it plans to deal with AI-generated content.

In May 2024, Facebook will begin applying a "Made with AI" label to content it can reliably detect as synthetic.

But the devil is in the details. How accurate will the detection models be? What AI-generated content will slip through? What content will be inappropriately flagged? And what will the public make of such labels?

While our work focused on Facebook spam and scams, there are broader implications.

Reporters have written about AI-generated videos targeting kids on YouTube and influencers on TikTok who use generative AI to turn a profit.

Social media platforms will have to reckon with how to treat AI-generated content; it's certainly possible that user engagement will wane if online worlds become filled with artificially generated posts, images and videos.

Shrimp Jesus may be an obvious fake. But the challenge of assessing what's real is only heating up.

Renee DiResta, Research Manager of the Stanford Internet Observatory, *Stanford University*; Abhiram Reddy, Research Assistant at the Center for Security and Emerging Technology, *Georgetown University*, and Josh A. Goldstein, Research Fellow at the Center for Security and Emerging Technology, *Georgetown University*