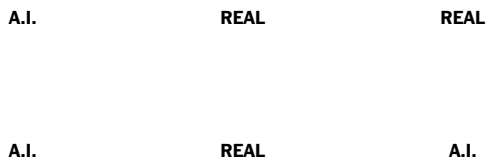


How Easy Is It to Fool A.I.-Detection Tools?

By Stuart A. Thompson and Tiffany Hsu June 28, 2023

The pope did not wear Balenciaga. And filmmakers did not fake the moon landing. In recent months, however, startlingly lifelike images of these scenes created by artificial intelligence have spread virally online, threatening society's ability to separate fact from fiction.

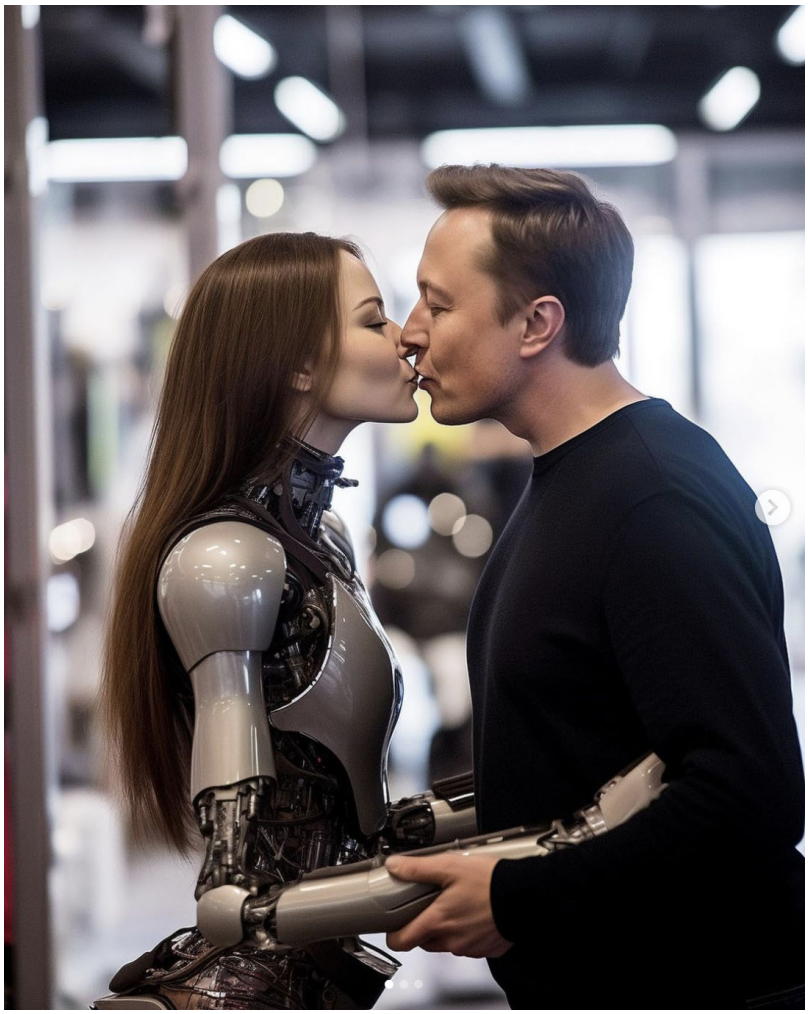
To sort through the confusion, a fast-burgeoning crop of companies now offer services to detect what is real and what isn't.

Their tools analyze content using sophisticated algorithms, picking up on subtle signals to distinguish the images made with computers from the ones produced by human photographers and artists. But some tech leaders and misinformation experts have expressed concern that advances in A.I. will always stay a step ahead of the tools.

To assess the effectiveness of current A.I.-detection technology, The New York Times tested five new services using more than 100 synthetic images and real photos. The results show that the services are advancing rapidly, but at times fall short.

Consider this example:

GENERATED BY A.I.



This image appears to show the billionaire entrepreneur Elon Musk embracing a lifelike robot. The image was created using Midjourney, the A.I. image generator, by Guerrero Art, an artist who works with A.I. technology.

Despite the implausibility of the image, it managed to fool several A.I.-image detectors.

Test results from the image of Mr. Musk

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
"REAL"	"REAL"	"A.I."	"A.I."	"A.I."

The detectors, including versions that charge for access, such as Sensity, and free ones, such as Umm-maybe's A.I. Art Detector, are designed to detect difficult-to-spot markers embedded in A.I.-generated images. They look for unusual patterns in how the pixels are arranged, including in their sharpness and contrast. Those signals tend to be generated when A.I. programs create images.

But the detectors ignore all context clues, so they don't process the existence of a lifelike automaton in a photo with Mr. Musk as unlikely. That is one shortcoming of relying on the technology to detect fakes.

Several companies, including Sensity, Hive and Inholo, the company behind Illuminarty, did not dispute the results and said their systems were always improving to keep up with the latest advancements in A.I.-image generation. Hive added that its misclassifications may result when it analyzes lower-quality images. Umm-maybe and Optic, the company behind A.I. or Not, did not respond to requests for comment.

To conduct the tests, The Times gathered A.I. images from artists and researchers familiar with variations of generative tools such as Midjourney, Stable Diffusion and DALL-E, which can create realistic portraits of people and animals and lifelike portrayals of nature, real estate, food and more. The real images used came from The Times's photo archive.

Here are seven examples:

A selection of test results 1/7

This A.I.-generated artwork of a smiling nun was created by Victoriano Izquierdo, a data scientist and artist who works with A.I.

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
"A.I."	"REAL"	"A.I."	"A.I."	"A.I."

Note: Images cropped from their original size.

Detection technology has been heralded as one way to mitigate the harm from A.I. images.

A.I. experts like Chenhao Tan, an assistant professor of computer science at the University of Chicago and the director of its Chicago Human+AI research lab, are less convinced.

"In general I don't think they're great, and I'm not optimistic that they will be," he said. "In the short term, it is possible that they will be able to perform with some accuracy, but in the long run, anything special a human does with images, A.I. will be able to re-create as well, and it will be very difficult to distinguish the difference."

Most of the concern has been on lifelike portraits. Gov. Ron DeSantis of Florida, who is also a Republican candidate for president, was criticized after his campaign used A.I.-generated images in a post. Synthetically generated artwork that focuses on scenery has also caused confusion in political races.

Many of the companies behind A.I. detectors acknowledged that their tools were imperfect and warned of a technological arms race: The detectors must often play catch-up to A.I. systems that seem to be improving by the minute.

“Every time somebody builds a better generator, people build better discriminators, and then people use the better discriminator to build a better generator,” said Cynthia Rudin, a computer science and engineering professor at Duke University, where she is also the principal investigator at the Interpretable Machine Learning Lab. “The generators are designed to be able to fool a detector.”

Sometimes, the detectors fail even when an image is obviously fake.

Dan Lytle, an artist who works with A.I. and runs a TikTok account called The_AI_Experiment, asked Midjourney to create a vintage picture of a giant Neanderthal standing among normal men. It produced this aged portrait of a towering, Yeti-like beast next to a quaint couple.

Test results from the image of a giant

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
"REAL"	"REAL"	"REAL"	"REAL"	"REAL"

The wrong result from each service tested demonstrates one drawback with the current A.I. detectors: They tend to struggle with images that have been altered from their original output or are of low quality, according to Kevin Guo, a founder and the chief executive of Hive, an image-detection tool.

When A.I. generators like Midjourney create photorealistic artwork, they pack the image with millions of pixels, each containing clues about its origins. "But if you distort it, if you resize it, lower the resolution, all that stuff, by definition you're altering those pixels and that additional digital signal is going away," Mr. Guo said.

When Hive, for example, ran a higher-resolution version of the Yeti artwork, it correctly determined the image was A.I.-generated.

Such shortfalls can undermine the potential for A.I. detectors to become a weapon against fake content. As images go viral online, they are often copied, resaved, shrunken or cropped, obscuring the important signals that A.I. detectors rely on. A new tool from Adobe Photoshop, known as generative fill, uses A.I. to expand a photo beyond its borders. (When tested on a photograph that was expanded using generative fill, the technology confused most detection services.)

The unusual portrait below, which shows President Biden, has much better resolution. It was taken in Gettysburg, Pa., by Damon Winter, the photographer for The Times.

Many of the detectors correctly thought the portrait was genuine; but not all did.

Test results from a photograph of President Biden

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
"REAL"	"REAL"	"REAL"	"REAL"	"A.I."

Falsely labeling a genuine image as A.I.-generated is a significant risk with A.I. detectors. Sensity was able to correctly label most A.I. images as artificial. But the same tool incorrectly labeled many real photographs as A.I.-generated.

Those risks could extend to artists, who could be inaccurately accused of using A.I. tools in creating their artwork.

This Jackson Pollock painting, called “Convergence,” features the artist’s familiar, colorful paint splatters. Most – but not all – the A.I. detectors determined this was a real image and not an A.I.-generated replica.

Test results from a painting by Pollock

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
“A.I.”	“REAL”	“REAL”	“REAL”	“REAL”

Illuminarty’s creators said they wanted a detector capable of identifying fake artwork, like paintings and drawings.

In the tests, Illuminarty correctly assessed most real photos as authentic, but labeled only about half the A.I. images as artificial. The tool, creators said, has an intentionally cautious design to avoid falsely accusing artists of using A.I.

Illuminarty’s tool, along with most other detectors, correctly identified a similar image in the style of Pollock that was created by The New York Times using Midjourney.

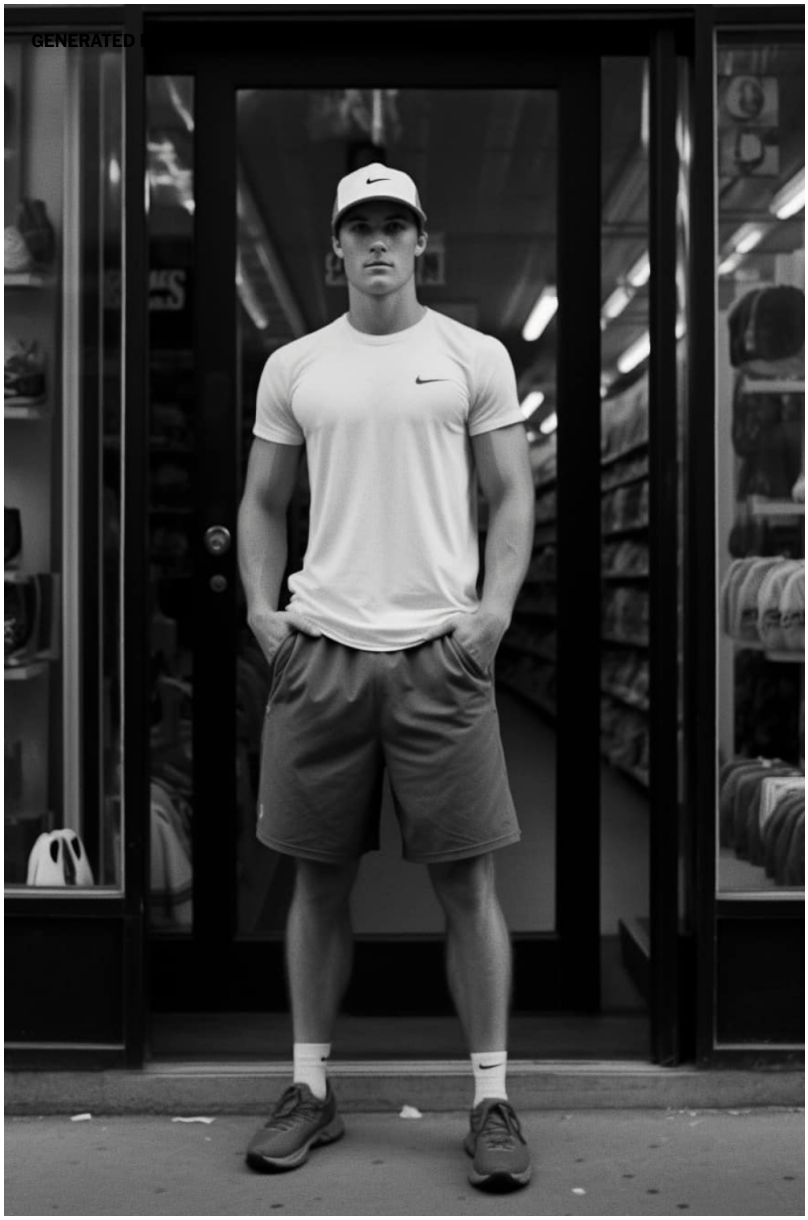
Test results from the image of a splatter painting

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
“REAL”	“A.I.”	“A.I.”	“A.I.”	“A.I.”

A.I.-detection companies say their services are designed to help promote transparency and accountability, helping to flag misinformation, fraud, nonconsensual pornography, artistic dishonesty and other abuses of the

technology. Industry experts warn that financial markets and voters could become vulnerable to A.I. trickery.

This image, in the style of a black-and-white portrait, is fairly convincing. It was created with Midjourney by Marc Fibbens, a New Zealand-based artist who works with A.I. Most of the A.I. detectors still managed to correctly identify it as fake.



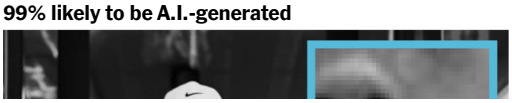
Test results from the image of a man wearing Nike

Umm-maybe	Illuminarty	A.I. or Not	Hive	Sensity
"A.I."	"REAL"	"A.I."	"A.I."	"A.I."

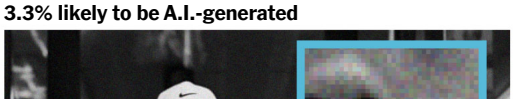
Yet the A.I. detectors struggled after just a bit of grain was introduced. Detectors like Hive suddenly believed the fake images were real photos.

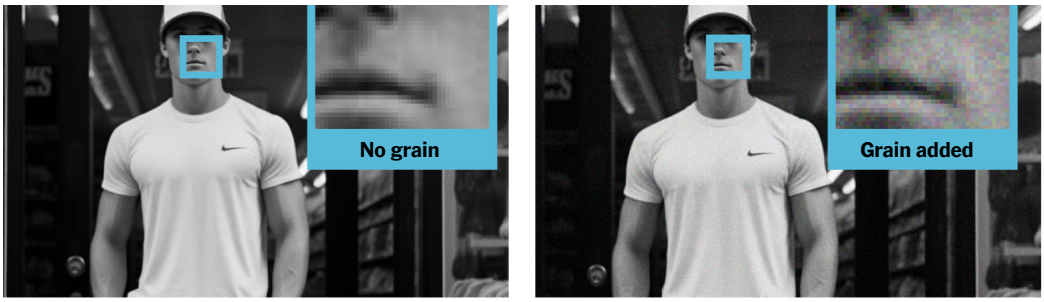
The subtle texture, which was nearly invisible to the naked eye, interfered with its ability to analyze the pixels for signs of A.I.-generated content. Some companies are now trying to identify the use of A.I. in images by evaluating perspective or the size of subjects' limbs, in addition to scrutinizing pixels.

99% likely to be A.I.-generated



3.3% likely to be A.I.-generated





Artificial intelligence is capable of generating more than realistic images – the technology is already creating text, audio and videos that have fooled professors, scammed consumers and been used in attempts to turn the tide of war.

A.I.-detection tools should not be the only defense, researchers said. Image creators should embed watermarks into their work, said S. Shyam Sundar, the director of the Center for Socially Responsible Artificial Intelligence at Pennsylvania State University. Websites could incorporate detection tools into their backends, he said, so that they can automatically identify A.I. images and serve them more carefully to users with warnings and limitations on how they are shared.

Images are especially powerful, Mr. Sundar said, because they "have that tendency to cause a visceral response. People are much more likely to believe their eyes."

Introduction photographs: Michelle V. Agins/The New York Times (barking dog); Tatiana Tsiguleva (A.I. house); Victoriano Izquierdo (A.I. nun); Absolutely AI (A.I. waves); Josh Haner/The New York Times (penguin); Ashley Gilbertson for The New York Times (creek); Damon Winter/ The New York Times (Trump poster; Gwyneth Paltrow; subway art); Julian van Dieken (A.I. close-up portrait); Holly Alvarez (A.I. children in library); Linus Ekenstam (A.I. fruit; A.I. men with helmets); Lam Yik Fei for The New York Times (man with face mask); Marc Fibbens (A.I. motorcycle).