‘Time is running out’

The search for new ways to spot undetectable deepfakes

Alex Hern
Technology editor

With more than 4,000 shares, 20,000 comments, and 100,000 reactions on Facebook, the photo of the elderly woman with her homemade 122nd birthday cake has unquestionably gone viral. “I started decorating cakes from five years old,” the caption reads, “and I can’t wait to grow my baking journey.”

The picture is unquestionably fake. If the curious candles – one seems to float in the air – or the weird amorphous blobs on the cake in the foreground didn’t give it away, then the fact the celebrator would be the oldest person in the world by almost five years should.

Thankfully, the stakes for viral supercentenarian cake decorators are low. Which is good, since as generative AI becomes better and better, the days of looking for telltale signs to spot a fake are nearly over. And that’s created a race against time: can we work out other ways to spot fakes, before the fakes become indistinguishable from reality?

“We’re running out of time of still being able to do manual detection,” said Mike Speirs, of the AI consultancy Faculty, where he leads the company’s work on counter-disinformation. “The models are developing at a speed and pace that is, well, incredible from a technical point of view, and quite alarming.

“There are all kinds of manual techniques to spot fake images, from misspelled words, to incongruously smooth or wrinkly skin. Hands are a classic one, and then eyes are also quite a good tell. But even today, it is time-consuming: it’s not something you can truly scale up. And time is running out – the models are getting better and better.”

Since 2021, OpenAI has released three versions of its image generator, Dall-E, each radically more capable than the previous. Indie competitor Midjourney has released six in the same period, while the free and open source Stable Diffusion model has hit its third version, and Google’s Gemini has joined the fracas. As the technology has become more powerful, it’s also become easier to use. The latest version of Dall-E is built into ChatGPT and Bing, while Google is offering its own tools for free to users.

Tech companies have started to react to the oncoming flood of generated media. The Coalition for Content Provenance and Authenticity, which includes among its membership the BBC, Google, Microsoft and Sony, has produced standards for watermarking and labelling, and in February OpenAI announced it would adopt them for Dall-E 3. Now, images generated by the tool have a visible label and machine-readable watermark. At the distribution end, Meta has started adding its own labels to AI-generated content and says it will remove unlabelled posts.

Those policies might help tackle some of the most viral forms of misinformation, such as in-jokes or satire that spreads outside its original context. But they can also create a false sense of security, says Speirs. “If the public get used to seeing AI-generated images with a watermark on it, does that mean they implicitly trust any without watermarking?”

That’s a problem, since labelling is by no means universal – nor is it likely to be. Fast companies such as OpenAI might agree to label their creations, but startups such as Midjourney don’t have the capacity to devote extra engineering time to the problem. And for “open source” projects such as Stable Diffusion, it’s impossible to force the watermark to be applied, since it’s always an option to simply “fork” the technology and build your own.

And seeing a watermark doesn’t necessarily have the effect one would want, says Henry Parker, the head of government affairs at factchecking group Logically. The company uses both manual and automatic methods to vet content, Parker says, but labelling can only go so far. “If you tell somebody they’re looking at a deepfake before they even watch it, the social psychology of watching that video is so powerful that they will still reference it as if it was fact. So the only thing you can do is ask: how can we reduce the amount of time this content is in circulation?”

Ultimately, that will require finding and removing AI-generated content automatically. But that’s hard, says Parker. “We’ve been trying for five years on this, and we’re quite honest about the fact that we got to about 70%, in terms of the accuracy we can achieve.” In the short term, this is an arms race between detection and creation: even image generators that have no malicious intent will want to try to beat the detectors since the ultimate goal is to create something as true to reality as a photo.

Logically thinks the answer is to look around the image, Parker says: “How do you actually try to look at the way that disinformation actors behave?” That means monitoring conversations around the web to detect malefactors in the planning stage on sites such as 4chan and Reddit, and keeping an eye on the swarming behaviour of suspicious accounts that have been co-opted by a state actor. Even then, the problem of false positives is difficult. “Am I looking at a campaign that Russia is running? Or am I looking at a bunch of Taylor Swift fans sharing information about concert tickets?”

Others are more optimistic. Ben Colman, the chief executive of image detection startup Reality Defender, thinks there will always be the possibility of detection, even if the conclusion is simply flagging something as possibly fake rather than ever reaching a definitive conclusion. Those signs can be anything from “a filter at higher frequencies indicating too much smoothness” to, for video content, the failure to render the invisible, but detectable, flushing that everyone shows each time their heart beats fresh blood around their face. “Things are gonna keep advancing on the fake side, but the real side is not changing,” Colman concludes. “We believe that we will get closer to a single model that is more evergreen.”

But tech, of course, is only part of the solution. If people really believe a photo of a 122-year-old woman with a cake she baked herself is real, then it isn’t going to take state-of-the-art image generators to trick them into believing other, more harmful things. But it’s a start.

> ***‘Things are gonna keep advancing on the fake side’***
>
> **Ben Colman**
> *CEO, Reality Defender*

Reality checks
How to tell fact from fabrication

AI-generated images are getting better and better, but there are still some telltale signs to check for.

Hands and limbs Most people have five fingers on each hand, two arms and two legs. Many AI generators get a bit more carried away. State-of-the-art technology is better at hands, but in any group scene, pay attention to figures in the background: there’s likely to be a surplus of legs, some odd hands, or an arm slung around a body-less shoulder.

Words Misspelled words, letters that blur together and mysterious characters can all be signs.

Hair Human hair is made of strands that flow from the head down. AI strands often have a less defined start and finish, and up close can look painted on. Watch out, though – normal image compression can also do funny things.

Symmetry In the real world, objects often come in pairs or groups. Think about earrings or cutlery: it’s an unusual situation to find mismatches. But some AI systems can forget what’s happening on the left side of a face once it comes time to render the right.

Textures Repeated patterns, fabrics and textures are notoriously difficult to render. In the real world, bricks tend to be a uniform size and shape across an entire building, while the floral print on a wallpaper will be stroke-for-stroke identical each time it repeats. Helpfully, this is one sign where a small variation is more likely to be fake than a large one: botched needlework might lead to a skewwhiff pattern on a dress, but it is unlikely to result in a gingham print being different at the top of the leg from the bottom.

Geometry Look at the space an image is in, and the objects within it. Are right angles right-angled? Does a wall seamlessly become part of a bookshelf in the background? Can you visualise how the sofa fits behind the table that looks flush to the wall? If not, the picture might have been created by a system that has no understanding of 3D space.

Consistency Are there multiple images purporting to show the same thing? Compare them! Generating multiple images of the same space from different angles and at different times is trivial in the real world, and absolutely cutting edge in AI. Even video generators such as Sora, which can create videos moving throughout a virtual space, will rarely pan back to show something they have panned away from, because doing so reveals they have “forgotten” what was there originally.

Don’t get hung up on AI If an image seems questionable enough that you’re poring over it, take a step back and consider whether you should trust your gut. Perhaps the image isn’t AI-generated at all – but it could still be the result of an AI face swap, edited on Photoshop the old-fashioned way, staged entirely, or even simply miscaptioned (a so-called “cheapfake”). **Alex Hern**

Faceswaps and padded popes
Deepfakes that rocked the internet

Dan Milmo
Alex Hern

Concern about doctored or manipulative media is always high around election cycles, but 2024 will be different for two reasons: deepfakes made by artificial intelligence and the sheer number of elections taking place.

About half the world’s population will vote this year – including India, the US, the EU and most probably the UK – and the technology could be highly disruptive. Here is a guide to some of the most effective deepfakes in recent years, including the first attempts to create hoax images.

▼ *An AI-generated deepfake image of Pope Francis wearing a luxury padded jacket went viral last year*

ILLUSTRATION: AI-GENERATED

▲ *Early pictures created by OpenAI's text-to-image model Dall-E. It was given the prompt 'an armchair in the shape of an avocado/ an armchair imitating an avocado' and told to create an image*

ILLUSTRATION: AI-GENERATED

The pope in a padded jacket, 2023
An image of Pope Francis apparently clad in a Balenciaga padded jacket was a landmark moment in generative AI and deepfakery. Created by the Midjourney image-making tool, it soon went viral because of its staggering level of realism.

"The pope image showed us what generative AI is capable of and how quickly this content can spread online," says Hany Farid, a professor at the University of California in Berkeley and a specialist in deepfake detection.

While the pope image was shared because of its combination of realism and the absurd, it underlined the creative power of widely accessible AI systems.

DeepDream's banana, 2015
The banana where it all began. In 2015, Google published a blogpost on what it called "inceptionism", but which rapidly came to be known as "DeepDream". In it, engineers from the company's photo team asked: what happens if you take the AI systems that Google had developed to label images – known as neural networks – and ask them to create images instead?

"Neural networks that were trained to discriminate between different kinds of images have quite a bit of the information needed to *generate* images too," wrote the team. The resulting hallucinatory dreamscapes were hardly high-fidelity, but they showed the promise of the approach.

Celebrity face swaps, 2017
Generating images entirely from scratch is hard. But using AI to make changes to existing photos and videos is slightly easier. In 2017, the technology was on the absolute cutting edge, requiring a powerful computer, a ton of imagery to learn from, and the time and wherewithal to master tools that weren't user-friendly.

But one example fell squarely within those limits: face-swapping female celebrities into porn. By late 2017, such explicit clips were being made, and traded, at a remarkable rate, initially on Reddit and then, as word spread and anger rose, on shadier and more hidden forums.

These days, according to a briefing note from the thinktank Labour Together, "one in every three deepfake tools allow users to create deepfake pornography in under 25 minutes and at no cost".

▲ *A faked image of Donald Trump with black voters. Joe Biden is also a popular subject of deepfakes*

ILLUSTRATION: AI-GENERATED

Jordan Peele/Obama video, 2018
Where pornography led, the rest of the media followed, and midway through 2018, face-swapping tools had improved to the extent that they could be used as a creative tool in their own right. In one such video, created by BuzzFeed, the actor and director Jordan Peele delivered an impression of Barack Obama – swapped into actual footage of the president himself, ending with an exhortation to "stay woke, bitches".

Dall-E's avocado armchair, 2021
In 2021, OpenAI released Dall-E, and face-swapping became old news. The first major image generator, Dall-E offered the science-fiction promise of typing a phrase in, and getting a picture out. Sure, those pictures weren't particularly good, but they were images that had never existed before – not simply remixed versions of previous pics, but wholly new things.

The first version of Dall-E wasn't great at photorealism, with OpenAI's demo selection showing one vaguely realistic imageset, a number of photos of an eerily pupil-less cat. But for more figurative art such as the armchair pictured above, it already showed promise.

Zelenskiy orders his side to surrender, 2022-23
Within a month of Russia's invasion of Ukraine an amateurish deepfake emerged of President Volodymyr Zelenskiy calling on his soldiers to lay down their weapons and return to their families. It was poor quality but prompted the real Ukrainian president to hit back on his Instagram account, telling Russian soldiers to return home instead.

But then the deepfake Zelenskiy came back a year later and underlined how the technology had improved. This clip urged Ukrainian soldiers to surrender again and was more convincing.

NewsGuard, an organisation that tracks misinformation, said comparing the two videos shows how far the technology has advanced in a short space of time.

Trump with black voters, 2024
This year a faked image emerged of Donald Trump posing with a group black men on a door step. There is also an image of Trump posing with a group of black women. The former president, who will face Joe Biden in the 2024 presidential election, is a popular subject of deepfakes – as is his opponent.

In a blog that compiles political deepfakes, Berkeley's Farid says the image appears to be "an attempt to court black voters". Farid expresses a general concern that deepfakes will be "weaponised in politics" this year.

Joe Biden robocalls, 2024
There are numerous examples of the US president's image and voice being used in a manipulative manner. In January Biden's faked voice was used to encourage Democrats not to vote in a New Hampshire primary, even deploying the Biden-esque phrase "what a bunch of malarkey". Steve Kramer, a political operative, admitted he was behind the faked calls. Kramer was working for Biden's challenger Dean Phillips, whose supporters have experimented with AI technology by creating a short-lived Phillips voice bot. Phillips's campaign said the challenger had nothing to do with the call. Kramer has said he did it to flag the dangers of AI in elections.