‘Time is running out’: can a future of undetectable deepfakes be avoided?

Publication Date: 2024-04-08

Author: Alex Hern

Section: Technology

Tags: Deepfake, Artificial intelligence (AI), Computing, features

Article URL: https://www.theguardian.com/technology/2024/apr/08/time-is-running-out-can-a-future-of-undetectable-deepfakes-be-avoided

With more than 4,000 shares, 20,000 comments, and 100,000 reactions on Facebook, the photo of the elderly woman, sitting behind 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 also unquestionably fake. If the curious candles – one seems to float in the air, attached to nothing – or the weird amorphous blobs on the cake in the foreground didn’t give it away, then the fact the celebrant 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 tell-tale 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 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’s image generator, Dall-E, has released three versions, 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 posts that aren’t labelled. Those policies might help tackle some of the most viral forms of misinformation, like in-jokes or satire that spreads outside its original context. But they can also create a false sense of security, says Spiers. “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. Big companies like 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, like 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, 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, the issue 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 capture malefactors in the planning stage on sites like 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, 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.” 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. • Join Alex Hern for a Guardian Live online event about AI, deepfakes and elections, on Wednesday 24 April at 8pm BST. Book tickets here