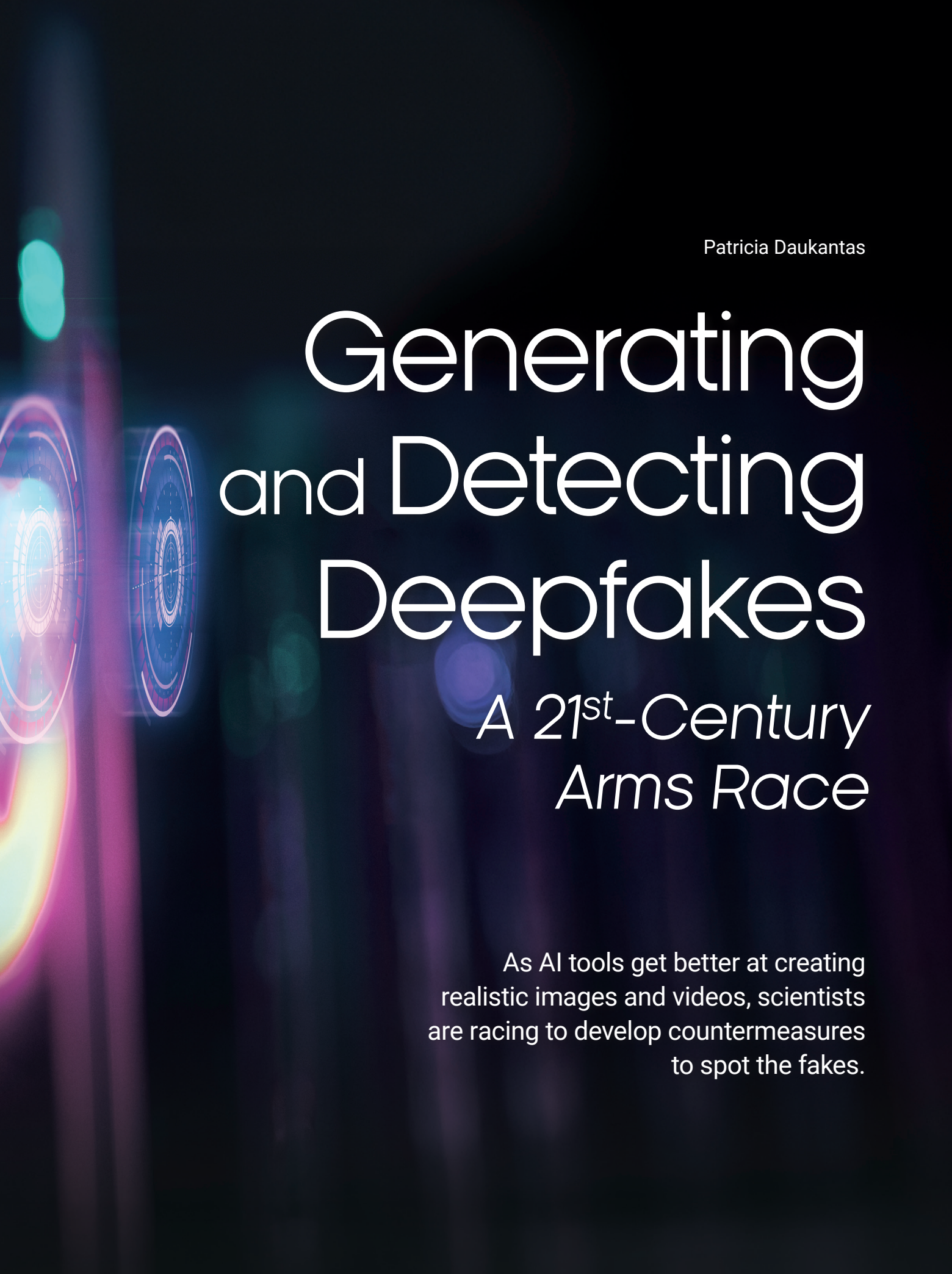
Getty Images

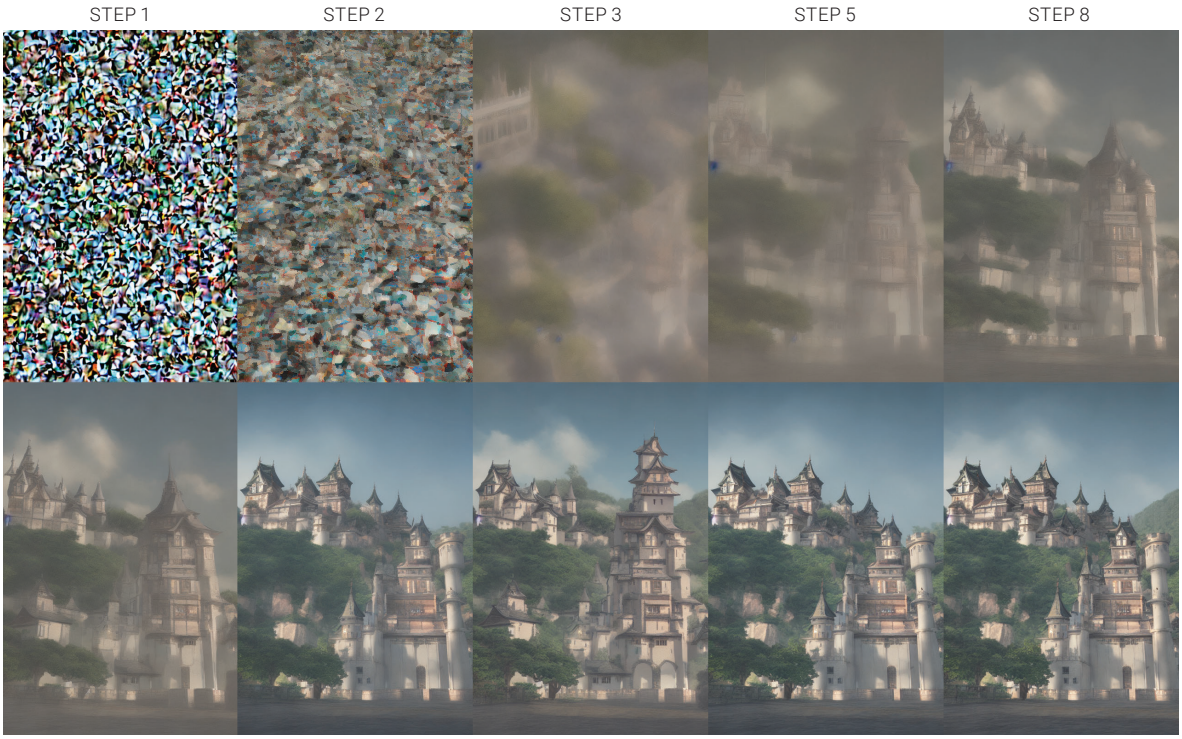


Patricia Daukantas

Generating and Detecting Deepfakes

A 21st-Century Arms Race

As AI tools get better at creating realistic images and videos, scientists are racing to develop countermeasures to spot the fakes.



The denoising process used by Stable Diffusion. The model generates images by iteratively denoising random noise until a required number of steps have been reached. The end result is an image—in this case, a European-style castle in Japan—that represents the concepts on which the text encoder was trained.

Benlisquare

Imagine this scenario: A finance worker signs onto a video call with several company officials. The firm's chief financial officer asks the employee to deposit millions of dollars into a specific bank account. Only after the transaction goes through does the truth come out: Everyone else on the video was a fake, and the money went straight to a fraudster.

Hong Kong police reported an investigation into just such a US$25 million scam in early 2024, hot on the heels of news stories about faked photos of celebrity singers and concerns over the spread of misinformation in election campaigns. The common thread among these worries: artificial intelligence (AI).

Since ChatGPT burst onto the scene in 2022, public awareness of AI has exploded. AI-fabricated images and videos have infiltrated society. One fake video shows the Philippine president purportedly taking drugs. Another depicts the Ukrainian president falsely inviting his soldiers to stop fighting against Russia.

Meanwhile, leaders in businesses from banks to dating apps worry about the future of identity verification in the wake of easily available software tools that can generate realistic forged renderings. The US Financial Crimes Enforcement Network (FinCEN) issued a broad alert warning financial institutions against fraud schemes powered by generative AI.

"The threat of deepfakes and synthetic media comes not from the technology used to create it, but from people's natural inclination to believe what they see," Dinusha Frings, chief executive officer of the European Association for Biometrics (EAB), said at an October 2024 workshop. "As a result, deepfakes and synthetic media do not need to be particularly advanced or believable in order to be effective in spreading this disinformation."

"There is an absolute need to try to identify what's real versus what's fake," says Kevin A. Pimbblet, director of the Centre of Excellence for Data Science, Artificial Intelligence and Modelling at the University of Hull, UK. "And it's quite the arms race." The first step to separating visual fact from fiction is understanding what goes into such manipulations, so that scientists can learn to identify telltale signs of deepfakes.

What are deepfakes, anyway?

Image manipulation is almost as old as photography itself. In 1860, an enterprising photographer "grafted" the head of incoming US president Abraham Lincoln

1047-6938/25/02/24/8-$15.00 ©Optica

Deepfakes go beyond mere pasting, retouching and photorealistic rendering. By definition, they have been created or manipulated by generative AI software.

onto the standing body of pro-slavery US senator and vice president John C. Calhoun to make it appear that Lincoln had posed for a heroic full-length portrait. (Calhoun could not object, as he had been dead for 10 years.) During the first half of the 20th century, Soviet leader Josef Stalin famously airbrushed his political opponents out of photographs.

Since the advent of raster graphics editors like Adobe Photoshop more than three decades ago, the media have published countless false or manipulated images of celebrities. And computer-generated imagery (CGI) made it possible to place fictional characters like Forrest Gump in the US Oval Office with John F. Kennedy.

Deepfakes go beyond mere pasting, retouching and photorealistic rendering. By definition, they have been created or manipulated by generative AI software. It's important to understand how the tools for generating deepfakes work in order to distinguish them from "real" images and films, or from the CGI that Hollywood employs.

Many machine-learning tools that create deepfakes fall into two categories: autoencoders and generative adversarial networks, or GANs.

An autoencoder is a type of artificial neural network that consists of an encoder, which maps a message to a code, and a decoder. Autoencoders are useful in areas beyond generating deepfakes: for example, they can be used to perform denoising in medical imaging routines, predict degradation in semiconductor lasers and aid optical communications.

The GAN is another deep-learning framework, developed about 10 years ago as an upgrade to a basic encoder. It pits the decoder (called the generator in this context) against another algorithm known as the discriminator. After the generator creates an image, the discriminator tries to determine whether the image is real or fake. The technology engages in a back-and-forth, zero-sum game that improves artificially generated images. It can also reconstruct astronomical images and produce detailed computational models in other branches of physics.

Autoencoders and GANs power a number of face-swapping and reenactment applications for creating deepfakes.

In face swaps, the manipulator pastes the facial characteristics of one person onto another human's body or hairstyle; deepfake generators incorporate 3D modeling methods so that the source images do not need to be in the same pose. According to Amit K. Roy-Chowdhury of the University of California, Riverside, USA, face swaps pose particular threats in scenarios where identity verification is important, such as video chatting or financial transactions.

With reenactment applications, one human subject can be made to mimic the expressions and movement of another. Altering the contexts of non-manipulated persons or objects is another way to confuse the viewer. A recent example showed a deepfake Pope Francis wearing an expensive puffer coat that he never donned in real life. Finally, an entire image or video can consist of people and things that never existed—a complete fantasy.

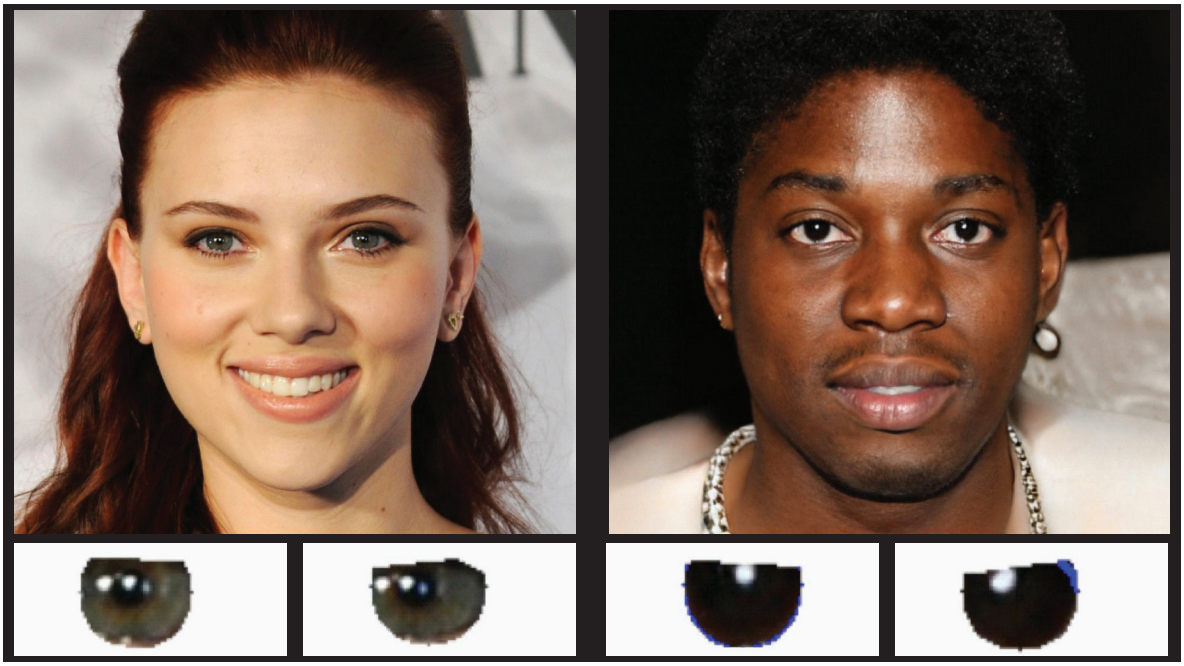
Recent advances in deepfake generation

According to Roy-Chowdhury, diffusion models have largely overtaken autoencoders and GANs for generating AI images. These models, including the Denoising Diffusion Probabilistic Model proposed in 2020, involve adding noise to an image, training a model to recreate the original from the noise, and then applying conditioning options related to specific images or textual inputs.

In the training process, the software keeps adding noise to the original image until it is pure noise, Roy-Chowdhury says. Then you train the machine-learning system, first so that it can recreate the image from the noise, and later to generate other images based on conditions added to the model. For example, the condition could be a text prompt describing a woman sitting on a chair.

The latest—and widely accessible—tools such as Midjourney, DALL-E and Stable Diffusion all use diffusion models "that are able to produce images that don't really exist, and they're increasingly quite convincing," Pimbblet says.

Some tools, which are now available online, even use AI to generate images based on text descriptions. Type "rainbow-colored unicorn" into the prompt box



In this image, the person on the left (actress Scarlett Johansson) is real, while the person on the right is AI-generated. Their eyeballs are depicted underneath their faces. The reflections in the eyeballs are consistent for the real person, but incorrect (from a physics point of view) for the fake person.

A. Owolabi

on one of these websites, and you'll get a portrait of a fantastical creature matching that description.

In the summer of 2023, US actors and screenwriters spent months on strike against major television and movie studios, as they sought protections against losing their livelihoods to AI. Later that year, a computer scientist and a student set out to create a deepfake of an American news anchor—just to prove it could be done.

With the cooperation of the CNN news-gathering network, Hany Farid of the University of California, Berkeley, USA, and a student, Matyas Bohacek of Stanford University, USA, copied clips of the anchor's voice from YouTube videos and passed them through a voice-cloning tool to produce an audio file of the person reading a fake script. Next, the pair fed the audio data and a cropped video of the anchor's head into an open-source tool, VideoRetalking, that changes the mouth on each frame to fit the new words. A GAN called CodeFormer deblurred the generated mouth and fixed missing teeth.

Taking the technology a step further, the pair next created an image of a fictional female TV anchor. Every step they took—making a still image, writing a short script, generating a voice to read the script and animating the still photo—involved AI tools. The result appears at https://github.com/matyasbohacek/AI-news-anchor.

Detection methods: what's fake, what's real

The first quarter of the 21st century has seen astounding advances in photorealistic imaging. In the mid-2000s, before the deepfake era took off, observers in a study involving 360 pairs of real and computer-generated images depicting the same subject correctly identified 83% of the real photographs and 82% of the virtual images. Around that time, Farid (then at Dartmouth College, USA) and Siwei Lyu, his then-student (now at the University of Buffalo, USA) developed a statistical model based on wavelet-like decomposition to distinguish between photographic and photorealistic images.

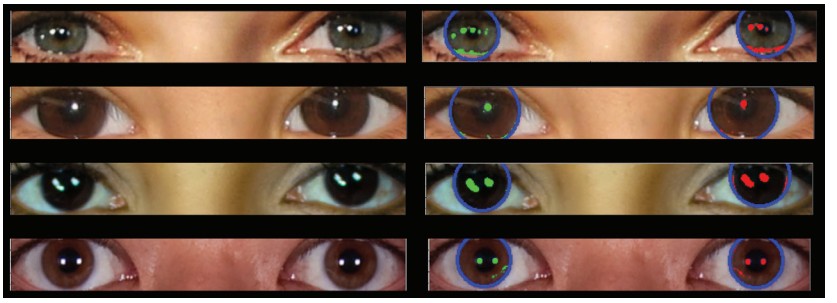
Some early attempts at deepfakes went viral for unintended reasons—because the artificial human had six toes or fingers, for example. The software behind the websites that make deepfakes for the casual end user—sites that have a limited number of templates or use older versions of algorithms—may still serve up odd-looking appendages or teeth. As tools have improved in the last few years, though, simple visual inspections may not be enough to ascertain the reality of an image.

Academic, corporate and government researchers have been simultaneously and vigorously pursuing deepfake detection methods. The US Defense Advanced Research Projects Agency (DARPA) led one of the major

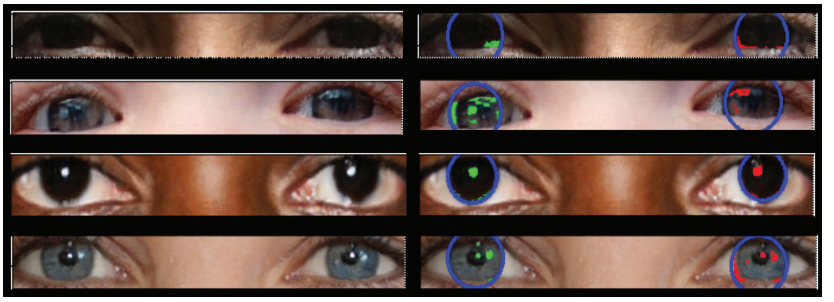
Academic, corporate and government researchers have been simultaneously and vigorously pursuing deepfake detection methods.

early programs, called Media Forensics, from 2016 to 2020. A successor DARPA program, Semantic Forensics (SemaFor), developed attribution algorithms to infer the provenance of potential deepfakes and characterization algorithms to assess whether they were created with malicious intent. SemaFor has published these algorithms in open-source repositories.

To train a machine-learning system to detect deepfakes, scientists provide examples of genuine images and deepfakes to the system and "tell" it which ones are which. Then the system teaches itself, not unlike the way a child learns by repetition how to recognize shapes and colors. After the system finishes with the training and validation datasets, researchers give it a known test dataset to evaluate the accuracy of the code. Finally, when the system is presented with a completely unknown image or video, people can assess the probability.



A series of real eyes showing largely consistent reflections in both eyes. The consistent reflections are highlighted in green and red in the right-hand column.



A series of deepfake eyes showing inconsistent reflections in each eye. The inconsistencies are highlighted in red and green in the right-hand column.

A. Owolabi

Although the algorithms are trained on thousands or even hundreds of thousands of images, one of AI's shortcomings is that it often fails to provide an explanation for why a particular image is a deepfake. "That still remains one of the research challenges in learning-based approaches," Roy-Chowdhury says. Explainability is an important concept in AI because companies have to build user and stakeholder trust that their deepfake detection algorithms are accurate.

It's all in the (physics) details

Physics-based details like incorrect lighting or shadows can indicate deepfake images. However, since those characteristics are specific to each image, it's hard to extrapolate from them to generalize over large numbers of images. Roy-Chowdhury says he's not sure if learning-based systems are picking up other subtle cues, like variances in the color temperatures of original photographs that are the basis for some deepfakes.

Pimbblet and his Hull colleagues believe they've found a detection technique based on light reflection from human eyeballs that has potential for broad application.

From a certain distance, the specular reflections in each eye should appear similar to the observer. (Look carefully at the glints in a newscaster's eyes under the bright lights of a television studio.) "But what you find is that the fake images don't quite have the physics right," Pimbblet says. Today's AI tools don't consistently generate the same highlights in each eye on a face. One eye in a deepfake portrait might reflect a single bright light, while the other may reflect multiple light sources. Some researchers have tried detection by subtracting one eye's reflections from those of the other eye; a small remainder means the image is real, while a large remainder means it's a fake.

Pimbblet, an astronomer by training, and a Hull M.Sc. student, Adejumoke Owolabi, used two statistical tools to quantify reflections in eye images in the same

The author used her laptop's built-in camera and the "deepfake sandbox" created by the Civic AI Security Program to generate deepfake images of herself in unlikely scenarios. The image that contains her hands does not depict the correct number of fingers.
P. Daukantas and deepfake.civai.org.

way they would classify galaxies by their morphology. One is the Gini coefficient, a statistical measure of how unevenly light is distributed across a collection of pixels (or across a galaxy); the other, dubbed "CAS parameters," quantifies concentration, asymmetry and smoothness of extended astronomical objects.

"It's not going to capture every single incidence of a fake image because some of them are just that good that they can fool it with the reflections," Pimbblet says. "But we can capture a decent chunk of them."

Pimbblet and Owolabi, who has since graduated, presented their results at a 2024 National Astronomy Meeting (UK) session on the application of astronomy techniques to earthbound problems. Before submitting their ongoing work to a journal, Pimbblet and his colleagues are creating their own dataset of forward-facing portraits in which the subjects' eyes are clearly visible. In principle, light reflections from other parts of a face—like the forehead, nose and cheeks of a person with oily skin—could be used for this consistency test, along with skin textures and blemishes.

In another line of research, scientists are incorporating ideas from facial recognition and biometric applications into deepfake-video detection. One mechanism, known as presentation attack detection, seeks to determine whether a video was created from a live person, an elaborate mask, or another non-living spoof by looking for eye blinking, nodding and other subtle clues.

The above techniques are generally called "handcrafted" detection methods because they involve some preprocessing by hand before the images are sent to a classification algorithm. Deep-learning methods show promise for detecting subtle clues within carefully crafted deepfakes. These deep-learning strategies often involve convolutional neural networks, sometimes in combination with recurrent neural networks.

Other promising work explores the potential usefulness of optical flow fields. Optical flow, according to Irene Amerini, University of Florence, Italy, and her colleagues, is a vector field that is computed on two consecutive video frames to detect the apparent motion between them. A single video frame might look completely "normal," but unnatural or inconsistent movement patterns in a subject's motion is a giant warning signal that a video has been manipulated.

A side-by-side comparison of the optical flow fields of two frames from real and deepfake videos reveals that the motion vectors around a moving object, such as the mouth of the video's subject, are noisier in real videos than in their phony counterparts. The Florence team and other researchers have been training convolutional neural networks to measure optical flow fields in videos suspected of fakery.

Ultimately, people need to evaluate the veracity of images and videos not just with deepfake detection tools, but also by assessing the context in which they appear and their provenance, Roy-Chowdhury says. Deepfake researchers also agree that they need larger and more diverse datasets for training algorithms. The deep-learning tools can only be as accurate as the information they have been fed.

Ethical concerns about deepfakes

While movie studios might spend months doing motion-capture studies of live actors before rendering CGI characters, Bohacek and Farid created their female

Ultimately, people need to evaluate the veracity of images and videos not just with deepfake detection tools, but also by assessing the context in which they appear and their provenance.

news anchor out of whole cloth in just two days. That speed makes the repercussions of these advances in deepfake generation being used for nefarious reasons even more ominous.

Recent elections in the United Kingdom, India and the United States have sparked public debate over the possible role of deepfake images and videos in propaganda and disinformation. Certainly, public figures with a big digital footprint, such as politicians and newscasters, give potential image manipulators plenty of material to feed into GANs and diffusion models. But some already envision a world where nothing can be trusted because everyone and everything could be faked—an online dating profile, a public health announcement about the latest emerging virus or the picture accompanying a colleague's message asking you to transfer US$5 million to an offshore bank account.

Roy-Chowdhury acknowledges that, from a computer science perspective, there's a lot of hype about the damage AI can do. "Knowing what I know about [deepfake] technology, it's one of the things I am worried about, because it can spread misinformation," he says. "Deepfakes by themselves are not usually that dangerous," he explains, "but the actions they can lead to, the consequences that they can lead to, are dangerous, ultimately, because people can make very wrong interpretations of them. And that could lead to really bad consequences."

In an effort to mitigate this threat, a nonprofit group called TrueMedia.org this year unveiled a website that claims to vet social media posts for AI content. Founded by computer scientist Oren Etzioni, the site accepts social media posts and user-uploaded files and runs them through an aggregation of AI detectors. Another nonprofit, the Civic AI Security Program, offers an "AI sandbox" that takes a snapshot of the user and, within seconds, swaps the user's face into preset templates of people in prison, in an elevator or riding a unicorn.

"It's an arms race, right?" Pimbblet says. "You're going to see … generation and detection in tandem with each other as one fights against [the] other." AI-generated images and the algorithms that detect them may be locked in an infinite feedback loop.

Add to that equation people's ability to spot AI. In November 2024, the Coca-Cola Company drew international criticism for remaking its 1995 Christmas-themed television commercial entirely with AI tools such as Stable Diffusion, DALL-E, ChatGPT and its advertising agency's proprietary software. Social media influencers caught unrealistic details in the video—the truck wheels turning out of sync with the vehicle's speed, the panting dog whose body doesn't move. Others panned the commercial for its unnatural airbrushed appearance.

It was a good reminder not to discount the deepfake detection power of the human mind. OPN

Patricia Daukantas (patd@nasw.org) is a freelance science writer based in Lanham, MD, USA.

References and Resources

- M. Bohacek and H. Farid. "The making of an AI news anchor—and its implications," Proc. Natl. Acad. Sci. 121, e2315678121 (2004).
- S. Lyu and H. Farid. "How realistic is photorealistic?" IEEE Trans. Signal Process 53, 845 (2005).
- H. Farid. "Digital doctoring: can we trust photographs?" in *Deception: From Ancient Empires to Internet Dating*, ed. B. Harrington (Stanford University Press, 2009).
- I. Amarini et al. "Deepfake video detection through optical flow-based CNN," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019).
- S. Adee. "What are deepfakes and how are they created?" IEEE Spectrum, https://spectrum.ieee.org/what-is-deepfake (published 29 April 2020, updated 8 March 2024).
- L. Verdoliva. "Media forensics and deepfakes: an overview," IEEE J. Sel. Top. Signal Process. 14(5), 910 (2020).
- G. Mazaheri and A.K. Roy-Chowdhury. "Detection and localization of facial expression manipulations," IEEE Computer Society and Computer Vision Foundation Winter Conference on Applications of Computer Vision (2022).
- M.S. Rana et al. "Deepfake detection: a systematic literature review," IEEE Access 10, 25494 (2022).
- N. Savage. "Exposing deepfake imagery," SPIE Photonics Focus 4(5) (September/October 2023).
- H. Lee et al. "The tug-of-war between deepfake generation and detection," 41st International Conference on Machine Learning, Vienna, Austria (2024).