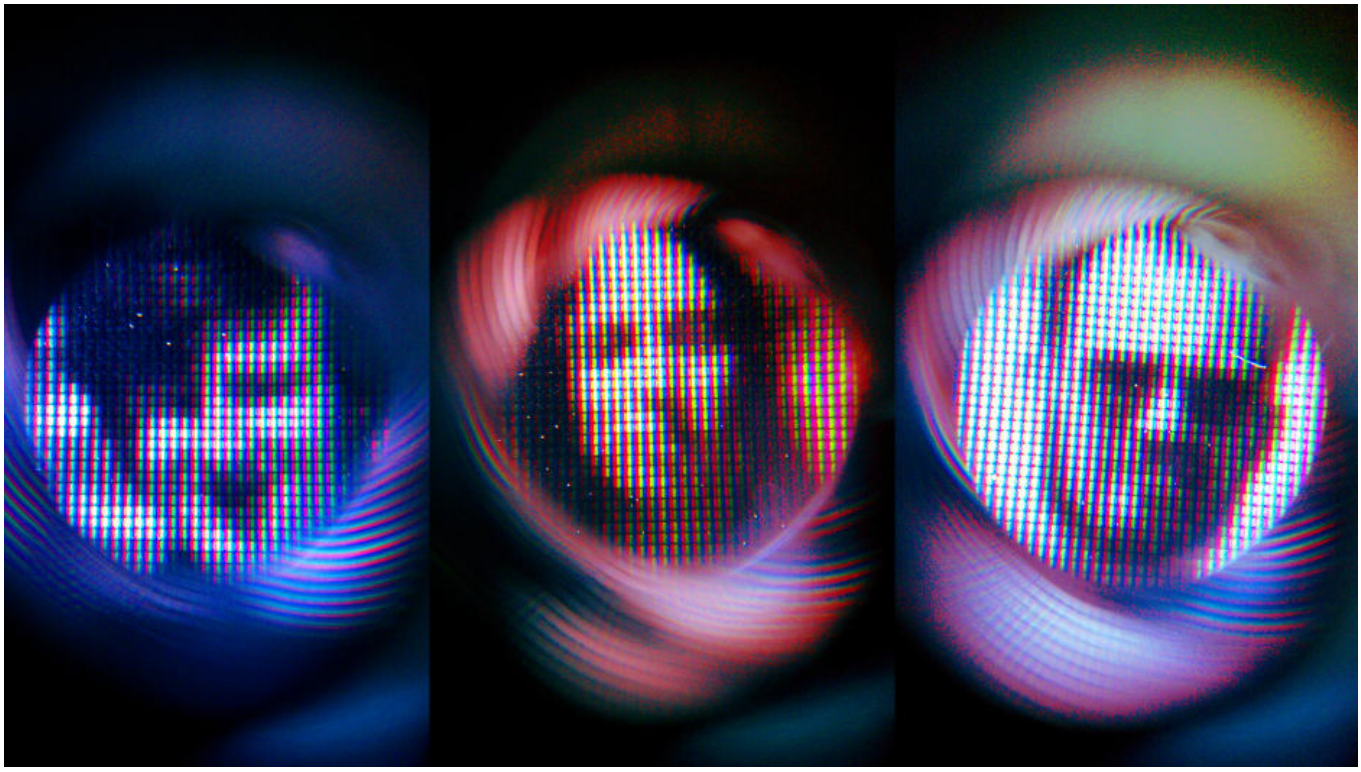
04.04.18 | ROBOT REVOLUTION

Can New Forensic Tech Win War On AI-Generated Fake Images?

As AI makes video manipulation easier, Defense Advanced Research Projects Agency scientists race to develop tools to detect what's real and what's not.



[Photos: Flickr user interestedbystandr]

BY STEVEN MELENDEZ

5 MINUTE READ

For gun lovers, the image was red meat: A Parkland school shooting survivor tearing the U.S. Constitution in two. It fed the NRA-fueled hysteria that, somehow, calling for tighter restrictions on assault weapons used in mass murders is a threat to the Second Amendment. The GIF went viral last week and conservatives went bonkers.

The problem: The animation, which looked pretty real, was fake. The teen March for Our Lives activist never put her hands on the Constitution—the animation was a doctored version of her shredding a shooting range target.

Welcome to the troubling world of "deep fakes." Earlier this year, Reddit took down a number of forums devoted to creating bogus videos, often pornographic, featuring one person's face swapped in for another. Open source artificial intelligence software makes the process radically simpler and more efficient than traditional video editing tools.

It's not just manipulation of visual media. Last year, Montreal startup Lyrebird released soundalike recordings of Presidents Barack Obama and Donald Trump, drawing attention to its speech synthesis technology. "They can make us say anything now," warned the faux Trump.

TAKING ON THE DEEP FAKES

Expert say these "deep fakes" are only going to get more sophisticated and common as tech tools for remixing photos, audio, and video get quickly perfected, developed, and made available to anyone with a laptop or smartphone. That creates a whole new set of problems for anyone trying to make it through April Fool's season and the midterm elections without falling for hoaxes and propaganda.

"I think it doesn't take a lot of imagination to see how this can go very badly very quickly," says Hany Farid, a computer science professor at Dartmouth College studying the issue.

It's also an issue for government and military officials trying to use third-party images to make decisions about events going on across the world.

"The central problem is that they don't want to use stuff that they're not sure of the provenance," says David Doermann, a program manager at the Defense Advance Research Projects Agency (DARPA). "They don't know where it came from, and they don't know if it's real or not."

DARPA is about a year and a half into a five-year program working with researchers around the country to build tools that can detect and analyze media manipulation. Right now, that's something that typically takes a lot of human analysis and intervention, says Farid, who is participating in the program.

Ideally, the tools would be able to flag altered images and explain just how they were altered—after all, a photo with dog ears spliced in by a smartphone filter but otherwise unchanged could still be useful to spies or reporters trying to understand a scene. The agency and its partners have put together infrastructure for researchers to test their algorithms on big datasets of real and fake media, getting results back within roughly a day, Doermann says.

Media manipulation isn't new, of course: Images were famously altered by the Soviet Union for propaganda purposes, and Adobe has been decrying the use of Photoshop as a verb since the early days of the web. Nor is the concept of spotting altered images new: Experts already have ways, many of them partly automated, to spot regions of a doctored photo that appear to be taken under different lighting conditions, with different cameras, or processed by different software. Researchers from Kitware, a Clifton Park, New York, company participating in the DARPA program, recently published papers on computationally detecting videos with dropped frames and spotting images where reflections don't match up with the rest of the scene.

"If you have a vehicle in that scene, and there's a puddle on the road in front of that vehicle, you should see potentially a reflection of that vehicle in that puddle," says Anthony Hoogs, Kitware's senior director of computer vision.

Many of the successful techniques have been published in scientific papers, so truly determined image manipulators can potentially circumvent them, but that requires a pretty sophisticated effort.

"Historically speaking, remember when fingerprints started to be used by police, 100 years ago—the thieves started to learn to use gloves," says Catalin Grigoras, director of the National Center for Media Forensics at the University of Colorado, Denver. "But in time, the forensic scientists came up with new developments, and usually in this kind of investigation, it's not about one [piece of] evidence only."

The difference now is that artificial intelligence technology, like the deep neural networks that have made speech parsing and facial recognition so commonplace, can now be used to quickly generate realistic-looking fake images.

FakeApp, a popular "deep fake" tool, is based on TensorFlow, the open source AI framework that grew out of Google's research. And some approaches now use what are called generative adversarial networks (GANs), an AI technique where one algorithm generates data while another algorithm attempts to tell real samples from fakes, to churn out increasingly realistic media.

"Eighteen months ago, nobody was thinking about this kind of machine-learning-generated content," says Farid. "It wasn't even on our radar screen 18 months ago when we started the whole [DARPA] program."

Image-generating GANs have plenty of applications—they can generate useful sample data to train other AI systems, like creating fake microscope images for algorithms to practice finding parts of the cell, for example, says Edward Delp, a Purdue University engineering professor participating in the DARPA program. But they're also a natural choice for generating media that can fool other algorithms, says Hoogs.

"The paradigm actually fits very well into defeating forensic detection techniques because you literally have a component in this learning system that is actively trying to fool a discriminator that's trying to decide whether the image is realistic or not," he says.

FAKE NEWS AND PROPAGANDA

And as those systems become more widespread and sophisticated, it's not hard to believe they'll find their way into the toolboxes of propaganda producers around the world. After all, Russian hackers are alleged to have widely circulated fake media during the 2016 U.S. election and other elections abroad, and hackers last year planted fake stories on the state-run Qatar News Agency's website with fake quotes from Qatar's emir, evidently to stir turmoil in the Middle East.

"One of the things that it may come down to is, unless you have some direct knowledge or believable information about the provenance of an image or a video, you may not believe, in the future, that anything you see is real," Delp says.

Still, participants in the DARPA program are optimistic that, when the program ends in a little more than three years, they'll have tools to detect and analyze fake images "automatically at internet scale," Farid says.

“I doubt we will have solved the problem in its entirety,” he says. “That is probably going to take a lot longer, at least another decade.”

ABOUT THE AUTHOR

Steven Melendez is an independent journalist living in New Orleans. More

Andy Cohen On How To Gracefully Secure A Bigger Role Within Your Company

These Architects Could Have The Solution To L.A.'s Homeless Crisis

How The Mayor Industry Should

IDEAS

IDEAS

Cape Town Isn't The Only Place That's Close To Running Out Of Water

IDEAS

This SimCity-Like Tool Lets Urban Planners See The Potential Impact Of Their Ideas

IDEAS

The Next Wave Of Tech-For-Good Companies Are Being Built By Women And Minorities

ENTERTAINMENT

ENTERTAINMENT

T.J. Miller arrested for fake bomb threat that has nothing to do with "Emoji Movie 2"

ENTERTAINMENT

Amid Fake News And Data Leaks, Ad Industry Makes Brand Safety Official

ENTERTAINMENT

Here's Why "Mean Girls" Crushes It As A Broadway Musical—And Why Most Movies Don't

CO.DESIGN

GRAPHICS

See Blockchain Rendered As A Mysterious Typeface

CITIES & SPACES

The Strange Beauty Of Brutalist Architecture, Mid-Demolition

INNOVATION BY DESIGN

Ikea's First Bluetooth Speaker Fits Perfectly Inside Your Ikea Furniture

FAST COMPANY

LEADERSHIP

Dear Blockchain Bros, I Won't Let Your Sexism Poison Our Industry

NEWS

Too much screen time? This eyewear startup may have a $95 solution

NEWS

Zuckerberg testimony live stream: How to watch the Facebook CEO's 2nd Congressional hearing