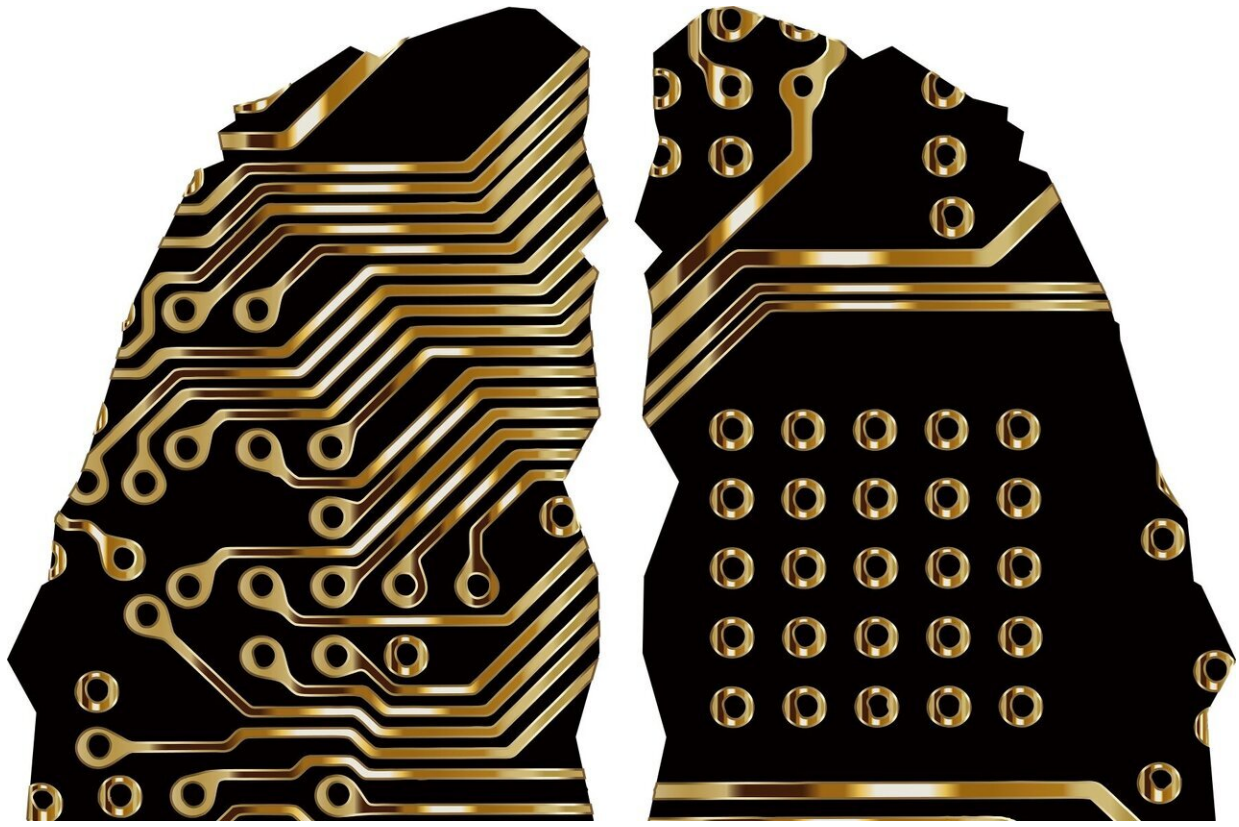
In a battle of AI versus AI, researchers are preparing for the coming wave of deepfake propaganda

October 9 2020, by John Sohrawardi and Matthew Wright



Credit: Pixabay/CC0 Public Domain

An investigative journalist receives a video from an anonymous whistleblower. It shows a candidate for president admitting to illegal

activity. But is this video real? If so, it would be huge news—the scoop of a lifetime—and could completely turn around the upcoming elections. But the journalist runs the video through a specialized tool, which tells her that the video isn't what it seems. In fact, it's a "deepfake," a video made using artificial intelligence with deep learning.

Journalists all over the world could soon be using a tool like this. In a few years, a tool like this could even be used by everyone to root out fake content in their social media feeds.

As researchers who have been studying deepfake detection and developing a tool for journalists, we see a future for these tools. They won't solve all our problems, though, and they will be just one part of the arsenal in the broader fight against disinformation.

The problem with deepfakes

Most people know that you can't believe everything you see. Over the last couple of decades, savvy news consumers have gotten used to seeing images manipulated with photo-editing software. Videos, though, are another story. Hollywood directors can spend millions of dollars on special effects to make up a realistic scene. But using deepfakes, amateurs with a few thousand dollars of computer equipment and a few weeks to spend could make something almost as true to life.

Deepfakes make it possible to put people into movie scenes they were never in – think Tom Cruise playing Iron Man – which makes for entertaining videos. Unfortunately, it also makes it possible to create pornography without the consent of the people depicted. So far, those people, nearly all women, are the biggest victims when deepfake technology is misused.

Deepfakes can also be used to create videos of political leaders saying

things they never said. The Belgian Socialist Party released a low-quality nondeepfake but still phony video of President Trump insulting Belgium, which got enough of a reaction to show the potential risks of higher-quality deepfakes.

Perhaps scariest of all, they can be used to create doubt about the content of real videos, by suggesting that they could be deepfakes.

Given these risks, it would be extremely valuable to be able to detect deepfakes and label them clearly. This would ensure that fake videos do not fool the public, and that real videos can be received as authentic.

Spotting fakes

Deepfake detection as a field of research was begun a little over three years ago. Early work focused on detecting visible problems in the videos, such as deepfakes that didn't blink. With time, however, the fakes have gotten better at mimicking real videos and become harder to spot for both people and detection tools.

There are two major categories of deepfake detection research. The first involves looking at the behavior of people in the videos. Suppose you have a lot of video of someone famous, such as President Obama. Artificial intelligence can use this video to learn his patterns, from his hand gestures to his pauses in speech. It can then watch a deepfake of him and notice where it does not match those patterns. This approach has the advantage of possibly working even if the video quality itself is essentially perfect.

Other researchers, including our team, have been focused on differences that all deepfakes have compared to real videos. Deepfake videos are often created by merging individually generated frames to form videos. Taking that into account, our team's methods extract the essential data

from the faces in individual frames of a video and then track them through sets of concurrent frames. This allows us to detect inconsistencies in the flow of the information from one frame to another. We use a similar approach for our fake audio detection system as well.

These subtle details are hard for people to see, but show how deepfakes are not quite perfect yet. Detectors like these can work for any person, not just a few world leaders. In the end, it may be that both types of deepfake detectors will be needed.

Recent detection systems perform very well on videos specifically gathered for evaluating the tools. Unfortunately, even the best models do poorly on videos found online. Improving these tools to be more robust and useful is the key next step.

Who should use deepfake detectors?

Ideally, a deepfake verification tool should be available to everyone. However, this technology is in the early stages of development. Researchers need to improve the tools and protect them against hackers before releasing them broadly.

At the same time, though, the tools to make deepfakes are available to anybody who wants to fool the public. Sitting on the sidelines is not an option. For our team, the right balance was to work with journalists, because they are the first line of defense against the spread of misinformation.

Before publishing stories, journalists need to verify the information. They already have tried-and-true methods, like checking with sources and getting more than one person to verify key facts. So by putting the tool into their hands, we give them more information, and we know that

they will not rely on the technology alone, given that it can make mistakes.

Can the detectors win the arms race?

It is encouraging to see teams from Facebook and Microsoft investing in technology to understand and detect deepfakes. This field needs more research to keep up with the speed of advances in deepfake technology.

Journalists and the social media platforms also need to figure out how best to warn people about deepfakes when they are detected. Research has shown that people remember the lie, but not the fact that it was a lie. Will the same be true for fake videos? Simply putting "Deepfake" in the title might not be enough to counter some kinds of disinformation.

Deepfakes are here to stay. Managing disinformation and protecting the public will be more challenging than ever as artificial intelligence gets more powerful. We are part of a growing research community that is taking on this threat, in which detection is just the first step.

Provided by The Conversation