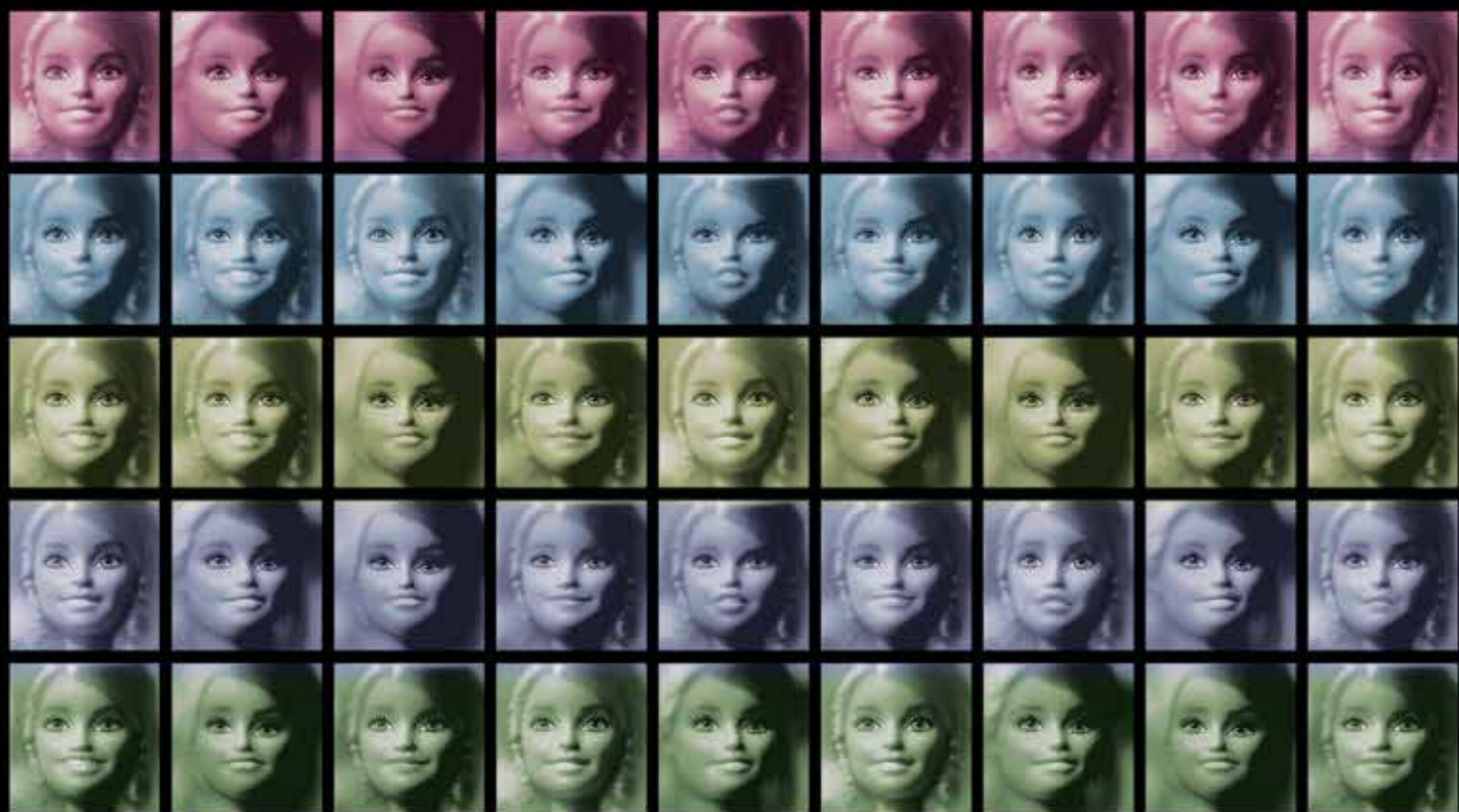



COMPUTING

DEEPPFAKE



Algorithm developed to detect images and videos altered by artificial intelligence, the new technological method for spreading disinformation

Sarah Schmidt

 In September, a doctored version of a clip from *Jornal Nacional*, the biggest news program on Brazil's Globo television network, was shared widely on social media. The video showed anchors William Bonner and Renata Vasconcellos announcing the results of a poll on voter intentions for the upcoming presidential election; however, the data on who was the preferred candidate was reversed, both in the graphics and in the words of the presenters. The next day, the show itself issued a warning that the video was a deepfake—where artificial intelligence (AI) is used to make highly convincing alterations—and was being used to misinform the population. This technology can be used to digitally imitate a person's face or simulate their voice, making them appear to do things they did not do or say things they did not say.

In August, another similarly altered video from the show, which once again inverted the results of a presidential election poll, was posted on TikTok, where it was viewed 2.5 million times according to the Comprova Project, a fact-checking group of journalists from 43 media outlets in Brazil.

"It could be deepfake technology that was used in these videos, but a more detailed analysis is needed. For us, what is important is knowing that they are fake," says computer scientist Anderson Rocha, director of the Computing Institute at the University of Campinas (UNICAMP), where he is head of the Artificial Intelligence Laboratory (Recod.ai). Rocha has been studying ways to detect malicious manipulation

of photos and videos—also known as synthetic media—including deepfakes.

In March 2022, shortly after Russia began its war against Ukraine, Ukrainian President Volodymyr Zelensky was the victim of a deepfake. A video circulated on social media in which he appeared to urge Ukrainians to lay down their weapons and return to their homes, suggesting that the country was surrendering. Facebook and YouTube removed the video as soon as it became apparent that it was fake. In the video, the president's face appeared on a near-motionless body wearing a green T-shirt.

When an original video is readily available for comparison, such as with the *Jornal Nacional* examples, it is fairly simple to verify that a video has been doctored. However, this is not always the case. Synthetic media is stripping the phrase "seeing is believing" of its meaning, and AI itself can be an ally to this process.

"Usually, synthetic videos are made in two stages. First, a deepfake platform is used to swap faces or synchronize mouth movements, and then they are edited using editing software," explains Rocha. Those who know what to look for can usually detect flaws in the program used to produce the fake video, such as inconsistent lighting or differences in contrast between the original video and the newly added face.

This process is similar to cutting a face out from one photo and sticking it onto another; the way the light falls and how the camera captures the two images are different in the two images. These small disparities serve as clues identifiable by computer forensics techniques, which is an area of research that has grown in recent years and with which Rocha is familiar.

Together with colleagues from the University of Hong Kong, Rocha has developed an algorithm that helps to detect whether the faces in a video have been manipulated and, if so, which regions have been changed. The program can determine, for example, whether the whole face has been doctored or just the mouth, eyes, or hair. “It was correct 88% of the time for low-resolution videos and 95% of the time for videos with a higher resolution,” explains Rocha, after the team tested the software on 112,000 faces, half of which were real and half of which were manipulated by four deepfake programs. It can also indicate whether an image was created from scratch rather than edited from an existing photograph. The results were published in the journal *Transactions on Information Forensics and Security* in April 2022.

According to computer scientists, other software has been developed that can detect evidence that videos are deepfakes; however, they mostly work by identifying clues left by well-known manipulation programs, which can be divided into two categories: those used to swap faces and those used to edit facial expressions. One platform is known to leave certain imperfections when synchronizing mouths; detection algorithms are then programmed to look for that specific error. “There is a problem with this. If we do not know what deepfake software was used, it becomes much more difficult to identify these traits. In addition, new applications are constantly being developed,” points out Rocha.

He and his colleagues thus trained their algorithm to detect clues without assuming any knowledge of the deepfake generator used. “We worked from the idea that regardless of the program, some noise will be left behind, something

that is not consistent with the rest of the image.” The software operates on two fronts. First, it looks for noise signatures, i.e., subtle changes around the edge of the face, for example; second, it determines a semantic signature, which can be a flaw in the color, texture, or shape.

“The algorithm automates the procedure a human expert would carry out, looking for inconsistencies, such as discrepancies in contrast,” he says. “The next step is to test it with fake videos generated by a larger number of programs to confirm its potential.”

This type of algorithm can be used for various purposes related to combating the malicious use of deepfakes. Rocha is part of an international program created by the US Department of Defense called Semantic Forensics, alongside researchers from the University of Siena and the Polytechnic University of Milan in Italy and the University of Notre Dame in the USA. The objective is to develop tools that automatically detect video and image manipulation. “We have already seen cases of doctored videos of military exercises in other countries that have multiplied the number of missiles to show greater military power,” he says.

These algorithms can also help identify political deepfakes, such as in the case of the Ukrainian president, or even pornographic deepfakes. It was the use of the technology in this area that garnered it fame at the end of 2017, when internet users began putting the faces of Hollywood celebrities onto the bodies of actors in pornographic movies. According to a September 2019 survey by the Dutch cybersecurity company DeepTrace Labs, 96% of deepfake videos online are nonconsensual pornography. Most victims are women, primarily actresses, but there have also been reports of cases involving people who are not famous. In July of this year, the Brazilian pop star Anitta was also the victim of a pornographic deepfake. The original video used in this attack had already previously been used to produce a deepfake with the face of actress Angelina Jolie.



In one fake video, Ukrainian President Volodymyr Zelensky appeared to urge his compatriots to lay down their weapons

A PERSON'S FACE OR VOICE CAN BE IMITATED, MAKING THEM APPEAR TO SAY THINGS THEY DID NOT ACTUALLY SAY

According to Cristina Tardáguila, program director at the International Center for Journalists (ICFJ) and founder of fact-checking specialists Agência Lupa, Brazil has already had to expose the truth behind several deepfakes. Programs that help detect synthetic media automatically can thus be valuable aids for journalists and fact-checkers who are

working against the clock. “In regard to misinformation, you have to respond quickly. It is important to invest in AI and tools that can help detect and identify this type of fake content as quickly as possible. That way, we can shorten the time between false content being shared and a check being made,” she explains.

“Deepfakes are the pinnacle of fake news. They can deceive people more easily because viewers believe they are watching something that truly happened. The audio can also be generated synthetically,” says journalist Magaly Prado, who is doing a postdoctoral fellowship at the Institute for Advanced Studies of the University of São Paulo (IEA-USP) and authored the book *Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação* (Fake news and artificial intelligence: The power of algorithms in the disinformation war), which was released by Edições 70 in July.

She emphasizes that despite being less well-remembered and less common, deepfake audio files can be spread easily on platforms such as WhatsApp, which is widely used by Brazilians. These files are made using a similar method as that used to make videos; with accessible software that keeps getting better, it is possible to simulate a person’s voice. The easiest victims are public figures, whose voices are easily available online. The technique can also be used for financial scams. “In one case, an employee of a technology company received a voice message from a top executive asking him to transfer some money to him. He was suspicious, and the message was analyzed by a security company, which verified that it was constructed using artificial intelligence,” Prado says.

Bruno Sartori, the director of FaceFactory, explains that producing well-made deepfakes, whether audio or video, is not simple—yet. His company creates synthetic media for commercial use and provides content for comedy shows on the television channels Globo and SBT.

In 2021, he worked on a commercial for Samsung in which the presenter, Maísa, interacted with herself as a child. The latter was created using deepfake technology. The virtual girl is shown to dance, play, and throw a laptop in the air. On another occasion, he had to put an actor’s face on the body of a stunt double. “To train the AI well, you need a big database of images and audio of the person you want to imitate. Good programs that offer high-quality processing also need to have advanced settings; otherwise, there can be visible flaws on the face or, with audio files, a robotic sounding voice,” he explains.

Sartori does not believe that the manipulated *Jornal Nacional* videos with the altered poll data were altered using AI. “In my analysis, the creators used traditional editing techniques, cutting and reversing the order of the audio. This is known as a shallowfake. However, if it is done well, it has just as much potential to deceive people,” he stresses. He notes that these programs will probably become lighter, smarter, and more accessible over the coming years.

There are some ways in which a person can protect themselves from misinformation created with the aid of technology. One is to pay attention to the fair use and privacy terms of the many free apps used in everyday life—from those that ask for access to a user’s photos to add fun effects to those that can store recordings of a user’s voice. According to UNICAMP’s Rocha, many apps store a large amount of data that can be shared for other purposes, such as training deepfake software.

Another important point is media awareness. “While software can help us highlight fake media, the first step is to be suspicious of everything we receive on social networks. In addition, check the sources of this information, research them,” he concludes. ■

Project

Déjà vu: Coherence of the time, space, and characteristics of heterogeneous data for integrity analysis and interpretation (no. 17/12646-3); Grant Mechanism Thematic Project; Principal Investigator Anderson Rocha; Investment R\$1,912,168.25.

Scientific article

KONG, C. *et al.* Detect and locate: Exposing face manipulation by semantic- and noise-level telltales. *Transactions on Information Forensics and Security*. Vol. 17. Apr. 2022.

Book

PRADO, M. *Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação*. São Paulo: Edições 70, 2022.