Deepfake

Marcin Wilkowski

In 2019 a video recording was posted online showing the founder of Facebook Mark Zuckerberg speaking. He warned against a situation where one person has complete control over the data and secrets of billions of people. The real issue is that Zuckerberg never said anything like that.

The recording takes a dozen or so seconds. Zuckerberg is sitting at his desk wearing a beige T-shirt; it is well known that, like other Silicon Valley billionaires, he avoids wearing suits and expensive clothes. Using verbal emphasis and hand gestures while speaking, Zuckerberg highlights key words: 'stolen data', 'control' and 'secrets'. The recording, however, was not a testimony to the revolution in thinking about privacy, which Facebook had turned into a commodity subject to ruthless commercialization. It was produced by the artists Bill Posters and Daniel Howe, members of the art collective Big Dada, using machine-learning methods. The computer program based on those methods used the original CBSN recording from 2017, which served as a matrix for the manipulated clip. Zuckerberg's face was subjected to automatic processing – the algorithm adjusted lip movements and facial expressions to the words of this false statement.

Big Dada / Public Faces (2019) [accessed 12.03.2021]. Available on Bill Posters: http://billposters.ch/projects/big-dada/

↗ An approximately three-minute-long film titled Big Dada/Public Faces (2019) featuring Marcel Duchamp, Marina Abramović, Mark Zuckerberg, Kim Kardashian, Morgan Freeman and Donald Trump through applying deepfake technology.

The artistic action of the Big Dada collective emphasized the growing threat from recordings manipulated by machine-learning algorithms.

Manipulation in film recordings is nothing new, after all achieving desired effects by using appropriate shots or cutting out frames had been practised long before film as a medium started to be recorded and edited digitally. However, IT solutions called 'deepfake' take visual manipulation to a new level.

Really deep manipulation

'Deepfake' is a phrase coined using two parts. 'Deep' is derived from 'deep learning', which, in a nutshell, makes it possible to generate new content on the basis of the sets used as a template, allowing an algorithm to learn. For example, millions of photographs of human faces can be used to train the algorithm in such a way that it is able to generate completely fictional images. Similarly, millions of texts become the basis for natural language-generation algorithms, and it is on the basis of such texts that the algorithm learns to create new content. Those solutions are successfully used, for example, by text-editing software that suggests words, and in applications that generate short information texts about the weather, accidents and the results of stock market trading sessions.

The effectiveness of such algorithms is measured by the greatest possible naturalness of artificially created content, regardless of whether it is a text with weather information, a photograph or a film showing a landscape panorama, which combines original recordings with advanced CGI (computer-generated imagery).

'Fake', in turn, emphasizes the objectives of using such solutions: in this case, the effect of using deep learning is to manipulate and mislead the audience. Its use in a feature film is not wrong as an element of the fictional creation of the presented world. When, however, it comes to a documentary film, a political speech or an event coverage, it may become a propaganda tool or cause political crises or conflicts.

Deepfake manipulation has been used, for example, against women, whose images available online were inserted into pornographic recordings – the victims have not only been stars and celebrities but also ordinary social-media users. The process of generating a few-minutes-long clip of that kind may take more than 40 hours, but it can be created on an ordinary home computer using free software, without the need to program anything.

As a method of manipulation, deepfake is therefore dangerous both due to the naturalness of the effects obtained and the logic of online communication. When we are constantly flooded with information, texts and images, there is no time or energy to analyse their authenticity. Sometimes, in the case of seemingly naturalized manipulation, it is impossible to recognize falsification in a few seconds. In addition, medicine, for example, is becoming the target of attacks using this method, in that image analysis is already the basis for recognizing certain diseases (such as lung cancer). In 2019 an attack carried out for research purposes at one of the medical centres in Israel allowed for the generation of deepfake lung scans by computer tomography. Fortunately, no one suffered because the break-in was revealed by the perpetrators themselves. During the 2016 election campaign, Hillary Clinton's physician made part of her medical records available online in response to allegations that the candidate's health did not permit her to take up the office of president of the USA. Had the medical records been manipulated, deepfake could therefore also have had an impact on political processes.

Deepfake in history

IT methods for generating deeply manipulated content are being developed in the scientific community as part of research projects aimed at increasing the possibilities of working with digital images. Sometimes they can also have original scientific applications, including researching the past. In archaeology, deep-machine analysis of satellite images supports the search for potential sites and the reconstruction of damaged objects, such as ceramics and coins.

In their 2016 publication, Shira Faigenbaum-Golovin and others demonstrated the usefulness of machine learning in analysing ancient Hebrew texts written on ceramic fragments around 600 BCE. They succeeded in proving that the texts were written by six different authors by automatically comparing letters from individual inscriptions.

In the historical sciences, deep learning is also successfully being used to recognize handwriting in medieval manuscripts and to complete Greek inscriptions. CBIR (content-based image retrieval) methods make it possible to search through huge collections of historical visual material. Instead of entering keywords as in traditional search engines, the user can search by image.

Machine learning also enables the extraction of selected phrases in a large body of historical texts: for example, the automatic recognition of dates, places, people and activities described in digitized military reports and 'operations orders': this method is called NER (named-entity recognition). Italian researchers have used machine learning to recognize images of destroyed historical monuments in old film recordings, assuming that any new information about a non-existent object can be valuable and useful in realizing their virtual, three-dimensional reconstruction.

The solutions behind deepfake manipulations may serve to inspire a new look at audio-visual historical sources. A train arriving at La Ciotat station is shown in *Train Pulling into a Station*, one of the first silent films made by the Lumière brothers in the late 19th century. For historical reasons, even the best digitized versions of this material are of poor quality, but with the help of machine learning, the recording has been rewritten to 4K resolution, making its details more visible now. It was actually a very similar operation to that used in deepfake manipulation: the recording was supplemented with new pixels based on available artificial intelligence libraries. So this is a reinterpretation, rather than a neutral reconstruction, which is a fundamental difference for any historical study. Black-and-white historical photographs automatically coloured by machine-learning software should be treated in a similar way.

Andy Myers, *[Actual 4K Scan] The Arrival of a Train at La Ciotat Station*, Lumière Brothers, 1896, 2020 [accessed 09 April 2021]. Available on YouTube: https://www.youtube.com/watch?v=1FAj9fJQRZA

↑ The Lumière brothers recorded the arrival of a train at the La Ciotat station in 1896. In February 2020 a refreshed digitally processed version of the recording was published in an improved 4K-resolution quality.

A responsible approach to artificial intelligence also requires questions about the ethics of the actions taken. In the forthcoming film *Finding Jack* (directed by Anton Ernst and Tati Golykh), a reconstructed image of James Dean is to appear, with new elements playing into the actor's image reconstructed from old recordings. Are such projects crossing an acceptable line?

In the project run by the USC Shoah Foundation in Los Angeles, California, Holocaust witnesses have become a model on which interactive holograms are built: artificial intelligence mechanisms are used not only to revive the virtual character, generated by images from 52 cameras, but also to prepare answers to questions asked by the interlocutors.

This dynamic, seemingly natural dialogue, which constitutes the core of the project, is based on a list of thousands of questions and answers specially developed in

relation to the biographies of witnesses and the reality of the Shoah. Although its authors emphasize that the hologram is rooted in the authentic memories of an actual witness who had agreed to take part in the experiment, one may wonder whether there is room for some abuse, especially if the response mechanism proposes statements that greatly overinterpret the original ones.

USC Shoah Foundation, Nimrod 'Zigi' Ariav, New Dimensions in Testimony, 2016 [accessed 09 April 2021]. Available on the USC Shoah Foundation: https://sfi.usc.edu/gallery/nimrod-zigi-ariav-new-dimensions-testimony-nov-3-2016

← The Holocaust witness Nimrod 'Zigi' Ariav recording his testimony in Hebrew, Los Angeles, California, April 2016.

The fight against the threat of deepfake consists in improving the methods of recognizing it. It is possible to analyse small errors and inconsistencies in the arrangement of the backgrounds, character movements and facial expressions, and it is also now possible, at a programming level, to examine changes occurring between successive recording frames and discover inconsistencies that suggest manipulation. Falsifications can also be detected using statistical methods, referring for example to Benford's law. This law determines the probability of occurrence of certain first digits in series of numbers and has been used for decades, for example, in detecting tax fraud. It turns out that these regularities can also be found in social media data and computer file structures. A disruption of Benford's distribution may indicate that a graphic file or a video recording has been manipulated, even if its visual layer appears to be authentic. Perhaps all these methods and tools will soon become part of the methodology for historical research, supporting work with digital sources that do not have their original analogue form and document important events in recent history. Surely, however, a technical analysis of this type of resource should be preceded by

its critical evaluation: checking the origin of the material, the context in which it was created and its authorship – such an approach can be useful for any material distributed independently through social media.

Translation: Mikołaj Sekrecki

Copyediting & Proofreading: Caroline Brooke Johnson