When seeing is disbelieving

Mark Coeckelbergh

The emergence of deepfake technology, with its potential to spread confusion and misinfor-mation, means it's time to develop a new level of critical spectatorship

Developments in synthetic image generation and manipulation lead to spectacular results. With the help of artificial intelligence (AI), facial expressions can be transferred from one person to another, allowing the creation of so-called "deepfakes": false images or videos that appear to be completely genuine.

A well-known example is the video using the voice of the American actor Jordan Peele impersonating Barack Obama, in which the former US president appears to speak about the dangers of false information – but in reality a mix of readily available video software and an AI program has been used to transfer Peele's own mouth movements to Obama's face.[1] "Stay woke, bitches" the president appears to tell viewers at the end of the clip. It's surprisingly hard to tell that this realistic-looking piece of video is faked.

Facial expressions can be created. Lip movements can be made to follow a script or can be captured directly from another person speaking. To the unsuspecting viewer, everything looks natural. Consider this use of real-time face capture and reenactment, in which a program called Face2Face[2], developed at Stanford University, is applied in real time to video clips of various presidents – George W Bush, Putin, Trump and Obama – to morph the facial movements of a second person onto the face of each president.

These technologies undermine trust in the veracity of images – and, in the end, trust in society. They can harm individuals, for example by creating seemingly compromising video, but they also have the potential to erode social and political institutions. How can justice work if a judge can no longer believe his or her eyes? How can democracy work if fake images of politicians circulate? Damaging videos of candidates could be created to manipulate elections. Deepfakes could potentially even start a war, as Nitesh Saxena, research director of the University of Alabama at Birmingham's department of computer science, has warned.[3]

So, who can we now believe – assuming, that is, that we had some trust in the first place. "Read my lips" no longer works if those lips can be manipulated by AI technology. Faced with deepfakes in the media, how can we trust anything or anybody?

What can be done? Technologists suggest that we could get machines to do the work – they tend to outperform humans in the business of detecting forgeries. So let AI work against AI. Facebook and DARPA (the US agency that develops emerging technologies for military use) are experimenting with this. It is possible to add digital watermarks to images. We need such technologies, just as we need anti-virus software. But at the same time they make us more dependent on private companies or non-transparent government organisations. And it is not yet clear how effective they are, even if they are better than humans.

Technological solutions, then, are not enough. Humans should also become more critical and be properly equipped with a new kind of scepticism, applied to this new medium. We have already got used to being healthily sceptical about older, established media. We don't unquestioningly believe everything we read. The camera never lies? That was being questioned when photography was still young, in the 19th century. The digital age just made it easier to deceive the eye. Photoshop and social media have shown us how simple it is to manipulate images: so-called "cheap fakes". The result is that today we do not immediately trust an image.

This ought to make us feel somewhat optimistic. In response to deepfakes, we should further develop this scepticism. Just as we educate children not to accept everything they read as gospel, we could also train them to be savvy when it comes to the images and video on their devices. Image scepticism, a new kind of critical spectatorship hygiene, would seem to be in order.

Professionals can help with this. Journalists need to train to better detect faked images. Is there distortion, is everything in sync, and are proportions correct, for example. This will become more difficult when AI technology improves; but we can leave that to AI. Instead, humans should focus on the history and context: everything that is said and shown around the image may indicate that something is fake, and from a source that should not be trusted. Who created the video, where, and for what purpose? Good journalism and good fact-checking are more important than ever.

Yet given that many people now derive most of their news from social media, there is often no journalist mediating. The only indication to go by is the source – for example, a traditional medium. But there is so much other content that is of interest to people, and often it is difficult to determine the source. How can individuals deal with this? How can platforms such as Facebook address the problem? The situation seems hopeless.

Technology has, of course, always been used to manipulate, deceive and undermine people, and to gain political power; in that sense, there is nothing new, even if that does not make it right. Illusion created by AI is fine if we all know that image manipulation has been employed. For example, it is now possible to create faces of people that do not exist, as the website This Person Does Not Exist[4] shows. Created by software engineer Philip Wang, the site creates an endless stream of fake portraits using a specially developed form of neural network. This is fine – interesting, even – as long as it's

clear that this is made by AI. It can be creepy or fun, but it's not morally wrong.

The problems start when we don't know that the technology has been applied. We need warning lights: tech solutions and a healthy scepticism based on common sense. Yet given these powerful technologies and the issues they raise, it may well take a while for this new common sense to develop. In the meantime, things are bound to get messy.