What if deepfakes made us doubt everything we see and hear?

Deepfakes are hyper-realistic media products created through artificial intelligence (AI) techniques that manipulate how people look and the things that they appear to say or do. They hit the headlines in 2018 with a deepfake video of Barack Obama, which was designed to raise awareness of their challenges. The accessibility and outputs of deepfake generation tools are improving rapidly, and their use is increasing exponentially. A wide range of malicious uses have been identified, including fraud, extortion and political disinformation. The impacts of such misuse can be financial, psychological and reputational. However, the most widespread use so far has been the production of non-consensual pornographic videos, with negative impacts that overwhelmingly affect women. Deepfakes may also contribute to worrying trends in our media, as well as in our social and democratic systems. While the technology itself is legal, some malicious uses are not, and a combination of legal and technical measures may be mobilised to limit their production and dissemination.

The name 'deepfake' combines 'deep' as in deep learning, and 'fake', as in manipulated or entirely fabricated. The best-known examples of deepfakes are videos that manipulate how people look, and the things that they appear to say or do. However, they can also include still images, audio or even written texts that are designed to present a distorted representation of events.

© fran_kie / AdobeStock.

In contrast to more traditional media manipulation techniques, deepfake production relies on an innovative deep learning technique called 'generative adversarial networks' (GANs), which can increase both the degree of automation and the quality of the output compared to conventional techniques. GANs generate deepfakes by pitting two AI agents – also described as artificial neural networks – against each other. While the producer agent learns to create fakes that look just like standard recordings, a detector agent learns to identify whether a media product is fake or authentic. A feedback loop is generated between the two so that, as the producer agent learns, it gets better at fooling the detector by creating more realistic fakes and, as the detector agent learns, it finds more sophisticated ways of identifying the fakes. In the end, the producer agent can create extremely realistic fakes and sometimes only its adversary – the detector agent – can tell that they are not authentic. An interesting side effect of this learning process is that the two agents improve together. So, by creating a great deepfake producer, you also create a great deepfake detector, and vice-versa.

Rudimentary use of deepfake production tools with limited resources may allow some generally low-quality results. However, producing high-quality deepfakes that really pass for authentic recordings requires substantial data and programming skills. Nonetheless, the increasing availability of data and accessible tools is making it easier for more people to make their own deepfakes.

Potential impacts and developments

The number of deepfake videos online is growing exponentially. The best-known examples have used the faces of famous people, such as Barack Obama and Tom Cruise, partly because there is so much data available. However, these examples were clearly labelled as deepfakes, and were designed to show the world what was possible, while highlighting opportunities and challenges. Legitimate applications

Author: Philip Boucher, Scientific Foresight Unit (STOA)

PE 690.046 - September 2021

include art, satire and entertainment, with notable examples including special effects and personal avatars.

Yet, the malicious use of deepfakes can also cause serious harm to individuals, as well as to our social and democratic systems. Deepfakes may be misused to commit fraud, extortion, bullying and intimidation, as well as to falsify evidence, manipulate public debates and destabilise political processes. Political disinformation is often cited as the biggest risk of deepfakes and, indeed, a well-timed deepfake during an election campaign could do enormous damage on several levels. Until now, however, the overwhelming majority of deepfakes have been pornographic videos produced without the consent of the women that are falsely depicted in them. This reveals a substantial gender discrimination aspect of the technology, because the negative impacts disproportionately affect women.

Those that are most directly affected by malicious deepfakes are the individual victims of fraud, blackmail, disinformation and non-consensual pornography. Targets have included citizens, businesses, and public figures. However, perhaps the biggest victim of deepfakes is the notion of truth. Just as manipulated videos can be presented as authentic, genuine recordings may also, as a result, be falsely dismissed as high-quality deepfakes. As such, simply knowing that deepfakes exist can be enough to undermine our confidence in all media representations, and make us doubt the authenticity of everything we see and hear online.

While manipulated media is nothing new, deepfakes may be more difficult to detect than previous techniques. Furthermore, various features of the current technical, social and legal context may enhance the risks associated with the technology. For example, the widespread use of social media and private messaging applications allows for the rapid dissemination and amplification of content with limited oversight. Our social context may also play a role, as deepfakes are well aligned with a growing climate of mistrust and polarisation. The legal status of deepfakes may vary across jurisdictions and could be further complicated by the possibility for malicious users to evade detection and enforcement efforts. Deepfakes are not the sole or even primary source of these social, technological and legal concerns, but they develop synergies with other malevolent elements of this context, benefitting from the environment to prosper while contributing to its maintenance and development.

Anticipatory policy-making

Deepfakes are, in themselves, perfectly legal, although some malicious applications are not. Some risks of malicious deepfakes may be mitigated through technical and legal measures, such as ensuring that they are properly labelled as non-authentic. The European Parliament has called for mandatory labelling of deepfakes, and this does indeed feature in the draft text of the proposed artificial intelligence act. The draft digital services act sets out rules for flagging and removing illegal content, which could help to interrupt their circulation and amplification. Both are currently under negotiation. In terms of technology, the EU's Horizon 2020 programme also supported the development of innovative responses to the challenges of deepfakes.

Such technical and legal measures cannot respond to all risks of malicious deepfakes, and their effectiveness will likely depend upon the technical and legal measures that are introduced to enforce them. Of course, any limitations need to be balanced against freedom of expression and freedom of the arts and sciences. However, while we are free to create media products such as deepfakes, we are not automatically entitled to have them widely circulated and seen. It is important to consider how malicious deepfakes are circulated and amplified online, as well as their role within wider social and political trends, because these are key factors in determining their resonance and impact. In this context, the European Parliament has stressed the importance of media pluralism, quality journalism and awareness-raising.

Read more about deepfakes and policy options in the STOA Study.