Spotlight on risk

Risk governance and the rise of deepfakes

Aengus Collins &
Touradj Ebrahimi

12 May 2021

Deepfakes first came to prominence less than five years ago. Since then, they have surged in quantity and quality, becoming both a source of viral entertainment and of concern about the dark side of digital life. In this article, we provide a risk governance perspective on the deepfake phenomenon, arguing that it warrants greater attention. We begin by distinguishing between three levels of harm that synthetic media can lead to: individual, organisational and societal. We then provide a simple framework for prioritising among these harms. Finally, we highlight the technical, legal and wider societal efforts that are under way to protect against deepfake risks.

In March 2021, a series of videos of Tom Cruise went viral on TikTok,[1] garnering 25 million views in three weeks. The scenes depicted are not remarkable, but the videos are: they are among the most convincing deepfakes ever produced. Deepfakes – videos and other digital content produced or manipulated using machine learning – first became prominent in 2017.[2] Since then, there have been significant improvements in their quality, and the technology for producing them has become increasingly accessible and user-friendly. The number of deepfake videos increased more than tenfold between 2018 and 2020.[3] Many deepfakes are, like the Tom Cruise examples, entertaining and ostensibly harmless. There are also examples of synthetic media

International Risk
Governance Center

being used for beneficial ends, such as in the entertainment industry or to create synthetic voices for people who have lost their speech.[4] But as manipulated digital content becomes more prevalent and realistic, the potential risks also increase.

Individual, organisational and societal risks

As the table below illustrates, there are three levels at which deepfakes can cause harm: individual, organisational and societal. The individuals harmed by deepfakes are almost always women, as pornographic videos account for the vast majority of documented deepfakes,[5] and the ability to swap a woman's face into a pornographic video make them a potential instrument of intimidation, coercion and abuse.[6]

For companies, the risks posed are less visceral, but potentially very costly. In 2019, an audio deepfake was used to persuade a CEO to transfer €220,000 to a fraudster's account.[7] More generally, any organisation that relies on documentary evidence – from courts[8] to insurance companies[9] – is potentially exposed to deepfakes.

The rise of deepfake technologies coincides with, and risks exacerbating, a wider set of societal problems relating to trust and truth in the information ecosystem. The idea that "seeing is believing" is a powerful one, and disrupting it could intensify patterns of disinformation that threaten informed decision-making in democratic societies. Deepfakes also sow doubt about authentic content. This can undermine trust among well-intentioned people, but it can also provide an excuse for dishonest actors to dismiss incriminating evidence as fake.[11]

A simple framework

One simple framework for policymakers to prioritise among these categories of deepfake risk is to consider the following three dimensions: severity (the level of harm caused by the deepfake), scale (how widespread the harm is) and resilience (the ability of the target to withstand the impact). This suggests a prima facie case for focusing on individual and societal risks. The impact of a deepfake on an individual is potentially severe and long-lasting, and many individuals may not have the resilience or resources to "bounce back" from an attack, particularly given the difficulty of having content removed from the internet.

The societal impact of deepfakes might cause a systemic deterioration – for example, due to a gradual erosion of trust – without having a sufficiently direct effect on enough people or organisations to trigger a response. However,

	Impact		
	Reputational damage	**Financial**	**Manipulation of decision-making**
Individual level	• Intimidation / abuse • Defamation	• Identity theft • Phishing-type scams • Extortion	• Attacks on politicians
Organizational level	• Brand damage • Undermining of trust in the organization	• Stock-price manipulation • Insurance fraud	• Fabricated court evidence • Media manipulation • Faked education papers • Attacks on political parties, advocacy groups, etc.
Societal level	• Damage to societal cohesion, norms of trust and truth, etc. • Domestic or foreign electoral manipulation • Deliberate stoking of tension / panic / conflict		

Source: Forged authenticity: Governing deepfake risks[10] the possibility of more dramatic societal impacts should not be discounted. A growing number of countries have witnessed increasing political polarisation and volatility in recent years, with digital misinformation playing a prominent role. Deepfakes could exacerbate this danger, and there is evidence of this already happening.[12]

Between the poles of individual and societal risks, the impact of deepfakes on organisations is likely to vary widely. Certain organisations could be badly damaged by a successful deepfake attack, but many will already have resources and processes in place — such as fraud-prevention teams — that could be adapted to respond to threats involving deepfakes.[13]

Technological responses: detection and provenance

There is no silver bullet for dealing with deepfakes. A mix of responses is needed, including technological, legal and broader societal measures.

In the area of technological responses, detection has been the main tool for combatting deepfakes, using algorithms trained to distinguish between authentic and fake content. The problem with this approach is that the arms race between deepfake generation and detection is unwinnable beyond the short term: new techniques for detecting deepfakes will be incorporated into the algorithms that generate deepfakes, leading to even more realistic output. The deepfake detection challenge organised by Facebook and a number of universities and technology companies highlights the difficulties.[14] Of the 35,000 algorithms submitted to the challenge, the winner had an accuracy rate of just 65%. Admittedly this does not reflect the cutting edge of detection technology, but even if an accuracy rate of 99.9% were one day achievable, the volume of content uploaded to the internet (for example, 720,000 hours of video to YouTube each day[15]) would still mean a huge number of deepfakes slipping through the net.

The scale of these challenges faced by detection technologies is such that tools for determining the origin, history and integrity of digital artefacts are attracting more attention as a way of establishing whether or not images or videos have been manipulated. One example of this is the Coalition for Content Provenance and Authenticity (C2PA),[16] led by Adobe, Microsoft and the BBC, which embeds additional metadata when digital content is created and certifies the source and history of the content. In parallel, the JPEG Committee, with the help of C2PA and other actors, is working to develop a universal standard with a similar objective: demonstrating securely and reliably the provenance of visual digital content, and indicating whether (and if so, how) it has been modified.

The motivation behind these provenance initiatives isn't to prevent manipulation; there are numerous good-faith reasons for altering content. The goal is to provide end-users with information about the status of any digital content they encounter. The idea is that this kind of transparency is a crucial underpinning if trust is to be restored in the digital ecosystem.

Legal and societal responses

As the prevalence of malicious deepfakes increases, we can expect increased legislative activity to help prevent and punish harm. In the US, there have already been a flurry of laws passed at state level,[17] and deepfake provisions were included in a 2019 federal defence law. Identifying and prosecuting the malicious use of deepfakes is difficult for numerous reasons, including jurisdictional barriers and the much greater ease of masking one's identity on the internet than in real life. Likewise, laws which seek to restrict or otherwise regulate content must be carefully balanced against those which protect freedom of expression. Nevertheless, there is a strong case for prohibiting deepfakes that are causing clearly demonstrable harm. Even if enforcement is difficult, laws play an important role in signalling societal boundaries.

In general, the legal status of deepfakes needs greater clarity, whether through new laws or improved guidance about existing laws. Progress is being made. For example, the EU's proposed new regulatory framework for artificial intelligence, published in April 2021, includes transparency obligations for systems designed to create deepfakes.[18] It is worth noting that internet platforms have also had to adapt to deepfakes. In 2020, Twitter introduced a new rule: "You may not deceptively promote synthetic or manipulated media that are likely to cause harm."[19]

Manipulated videos and other content are just one facet of radical changes in the information ecosystem over recent decades, which threaten to undermine the "epistemic security" of democratic societies.[20] Education to improve levels of digital literacy and critical thinking are crucial, but will not be a panacea. The volume and velocity of the digital information with which we are now confronted means it may be unrealistic to assess it effectively. The fact that viral content appears to appeal to emotional rather than rational drivers may also limit the effectiveness of increased critical engagement.[21]

Conclusion

In some senses, deepfakes are not particularly new. They are the latest iteration of age-old patterns of deception. However, the changes facilitated by digital technology are profound. The reproducibility, durability and global reach of digital content alters the potential scale and impact of deception. If an individual is targeted with a pornographic deepfake, that content is instantly available everywhere and could remain part of their digital legacy forever.

We should not overstate the risks posed by deepfakes. Producing high-quality fakes still requires skills and resources that most users do not have, and many of the harms remain possibilities rather than documented facts. But the quality and prevalence of deepfakes will keep increasing, and therefore so will the risks. Deepfakes of still images are now effectively indistinguishable from authentic images.[22] Videos are rapidly moving in the same direction. One area to monitor is audio deepfakes of people's voices. These are more difficult to create, but when their quality improves, it will be a potential game-changer, sowing confusion over telephone networks and allowing malicious actors to distribute convincing videos of anyone saying anything. It is a collective responsibility to ensure that machine learning developers, social media users and citizens more generally are educated about the harms that can result from the misuse of deepfake technologies.

↦ Read Forged authenticity: Governing deepfake risks

Acknowledgements
The authors would like to thank Martha Crawford, Kelsey Farish and Marie-Valentine Florin for reviewing this article.

[1] Tom (@deeptomcruise) TikTok. *TikTok*.
[2] Adee, S. What are deepfakes and how are they created? *IEEE Spectrum: Technology, Engineering, and Science News* (2020).
[3] Cavalli, F. How to detect a deepfake online with no coding skills. *Sensity* (2021).
[4] Ajder, H. The ethics of deepfakes aren't always black and white. *TNW Podium* (2019).
[5] Reports. Sensity.
[6] Hao, K. Deepfake porn is ruining women's lives. Now the law may finally ban it. *MIT Technology Review* (2021).
[7] Tung, L. Forget email: Scammers use CEO voice 'deepfakes' to con workers into wiring cash. *ZDNet* (2019).
[8] Maras, M.-H. & Alexandrou, A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *International Journal of Evidence and Proof* 23, 255–262 (2019).
[9] McMahon, L. Triple-I Blog. "Deepfakes": A looming nightmare for insurers? (2018).
[10] Collins, A. Forged authenticity: Governing deepfake risks. (2019).
[11] Gregory, S. Authoritarian regimes could exploit cries of 'deepfake'. *WIRED*. (2021).
[12] Cahlan, S. How misinformation helped spark an attempted coup in Gabon. *Washington Post* (2020).
[13] Viña, S. Digital deception: Is your business ready for "deepfakes"? (2020).
[14] Vincent, J. Facebook contest reveals deepfake detection is still an 'unsolved problem'. *The Verge*. (2020).
[15] Hale, J. More than 500 hours of content are now being uploaded to YouTube every minute. *Tubefilter* (2019).
[16] Overview C2PA.
[17] Ferraro, M. Deepfake legislation: A nationwide survey – State and federal lawmakers consider legislation to regulate manipulated media. (2019).
[18] European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2021).
[19] Our synthetic and manipulated media policy. Twitter Help.
[20] Seger, E. et al. Tackling threats to informed decision-making in democratic societies. (2020).
[21] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* 359, 1146–1151 (2018).
[22] This person does not exist.