CDEI Snapshot Series

Deepfakes and Audio-visual Disinformation

September 2019

Centre for Data Ethics and Innovation

About this Snapshot Paper

The Centre for Data Ethics and Innovation (CDEI) is an advisory body set up by the UK government and led by an independent board of experts. It is tasked with identifying the measures we need to take to maximise the benefits of AI and data-driven technology for our society and economy. The CDEI has a unique mandate to advise government on these issues, drawing on expertise and perspectives from across society.

The CDEI Snapshots are briefing papers that aim to improve public understanding of topical issues related to the development and deployment of AI. They are intended to separate fact from fiction, clarify what is known and unknown about an issue, and suggest questions for further investigation. Unless otherwise stated, the papers focus on present day or near-term issues. We do not intend to comment here on the long-term risks of AI and data-driven technology.

To develop this Snapshot Paper, we undertook a review of academic and grey literature, and spoke with the following experts:

— **Hany Farid**, Dartmouth College

— **Alissa Davies**, **Lynette Webb** and **Clement Wolf**, Google

— **Tim Hwang**, Harvard-MIT Ethics and Governance of AI Initiative

— **Alan Zucconi**, London College of Communication

— **Siwei Lyu**, University at Albany-SUNY

Contents

Summary

Reading time: 13–15 minutes

— Deepfakes can be defined as visual and audio content that has been manipulated using advanced software to change how a person, object or environment is presented.

— Deepfakes have become synonymous with **face replacement**, where someone's face is digitally mapped onto that of another person. Face-swapping, as this type of deepfake is also known, was first used in the doctoring of pornographic videos.

— Deepfakes can also take the form of:

  – **Face re-enactment**, where advanced software is used to manipulate the features of someone's face, with no face swapping involved

  – **Face generation**, where advanced software is used to create entirely new images of faces, which do not reflect a real person

  – **Speech synthesis**, where advanced software is used to create a model of someone's voice.

— Deepfakes are likely to become more sophisticated over time. For now, however, high quality content remains difficult to create, requiring specialist skills and professional software that is not yet widely available. We are yet to see a convincing deepfake of a politician that could distort public discourse.

— However, even rudimentary deepfakes can cause harm. Some estimate that there are now thousands of deepfake pornographic videos on the internet, which while clearly doctored still cause distress to those who have been featured.

— The growth of 'shallowfakes' is also concerning. This term refers to audio or video content that has been edited using basic techniques, such as where the captions have been changed or the footage slowed down to misrepresent events or the people featured.

— Legislation will not be enough to contain deepfakes or shallowfakes. We also need to invest in new screening technology that can check for irregularities in visual and audio content, and to educate the public about the existence of doctored material.

— Regardless of the measures taken to manage the spread of manipulated media, the government and regulators must be careful not to suppress benign and innovative uses of audio and visual manipulation, for example in the entertainment and marketing industries.

Disinformation, in particular 'fake news', has become a globally recognised phenomenon.[1] Think tanks, media pundits, parliamentary select committees and tech companies have all commented on its rise and recommended measures to contain it. For the most part, public concern has centred on fabricated text, including news articles, Facebook posts and Twitter accounts.[2] However, with the advent of 'deepfakes' – or visual and audio disinformation – some have speculated that we are entering a new chapter in the battle for truth on the internet.

Here we examine the nature of deepfakes, the risks they pose to society, and the potential measures that could oversee their use.

1 The UK government defines disinformation as '*the deliberate creation and dissemination of false and/or manipulated information that is intended to deceive and mislead audiences, either for the purposes of causing harm, or for political, personal or financial gain. 'Misinformation' refers to inadvertently spreading false information.'*

2 A recent controversy has been the launch of an impressive new language model (GPT2), which can be used to create lengthy passages of plausible text based on a source feed of a few sentences. This has sparked fears that text-based disinformation could become increasingly sophisticated in the years ahead.

1. What are deepfakes?

The term 'deepfakes' first became prominent in 2017 when a Reddit user, who went by the same name, began posting digitally altered pornographic videos on the website's forums.[3] These were doctored such that the faces of adult entertainers were replaced with those of other people, typically celebrities. Reddit soon became a focal point for the sharing of doctored porn videos, leading the tech media outlet Motherboard to run with a headline in 2017 that 'AI-assisted Fake Porn is Here and We're All F**ked'.[4]

For the purposes of this paper, deepfakes can be defined as visual and audio content that has been manipulated using advanced software to change how a person, object or environment is presented. The four main types are:

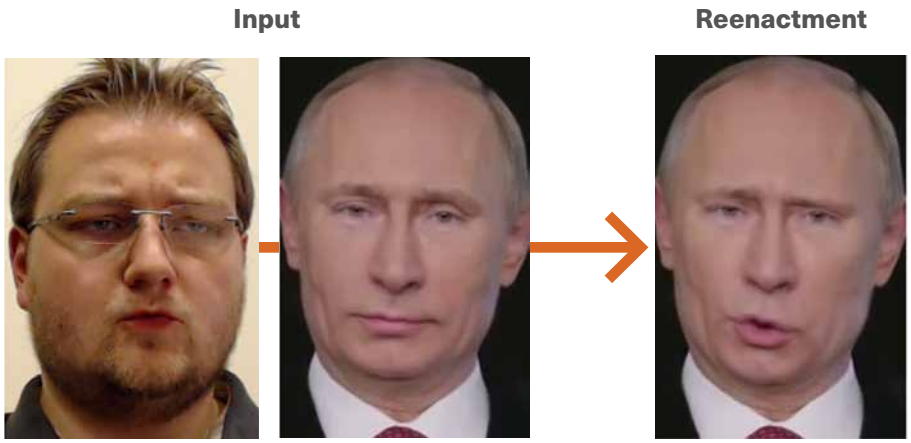
1. **Face replacement** – Otherwise known as face swapping, face replacement involves taking an image of someone's face (the source) and carefully 'stitching' it onto that of another person (the target). The identity of the target is concealed, with the focus being on the source.

Hillary Clinton's face is digitally 'stitched' onto Saturday Night Live actress Kate McKinnon.
Source/creator: Derpfakes

3 'Deepfakes' takes it name from the use of deep learning techniques to train visual manipulation algorithms.

4 Cole, S. (2017) *AI-Assisted Fake Porn is Here and We're All F**ked* [article] Motherboard, 11 December 017.

2. **Face re-enactment** – Also known as puppetry, face re-enactment entails manipulating the features of a target's face, including the movement of their mouth, eyebrows, eyes and the tilting of their head. Re-enactment does not aim to replace identities but rather to contort a person's expressions so they appear to be saying something they are not.



Researchers use the Face2Face tool to manipulate the facial expressions of Vladimir Putin.
Source/creator: The Visual Computing Lab at TUM

3. **Face generation** – Face generation involves creating entirely new images of faces. This is done using Generative Adversarial Networks, a novel form of deep learning that works by pitting two neural networks against one another: the first to generate an image, and the second to judge whether that output is realistic.[5]

Source/creator: ThisPersonDoesNotExist.com

5 ThisPersonDoesNotExist.com is a popular website that generates a novel face every time the page is refreshed.

4. **Speech synthesis** – A relatively new branch of deepfakes, speech synthesis involves creating a model of someone's voice, which can read out text in the same manner, intonation and cadence as the target person. Some speech synthesis products, such as Modulate.ai, allow users to choose a voice with any age and gender, rather than to emulate a specific target.

Each form of deepfake has its limitations. **Face re-enactment** keeps people's features intact and can therefore appear more life-like. But for now, targets must remain largely motionless and face head on to a camera (e.g. as a politician does while giving a public address). **Face replacement**, meanwhile, requires careful modelling and an understanding of how the source and target person look from multiple angles. Until speech synthesis matures, both face re-enactment and face replacement will require voice impersonators to provide the audio to run in parallel with videos. However, it can be difficult to marry sound with footage, as was demonstrated when Buzzfeed attempted to create a deepfake of Barack Obama.[6]

It is important to note the distinction between these four **forms** of deepfake and the **tools** used to create them. The latter range from Snapchat filters used to power rudimentary face swapping, to the Face2Face method used for face re-enactment. (Confusingly, 'deepfake' is also the name of a specific method for editing videos). As new tools emerge, deepfakes will improve in quality and become more varied. Although this paper only looks at the synthesis of faces and voices, any dimension of visual and audio content could in theory be manipulated, including background landscapes and the movement of entire bodies within video footage.[7]

6 Mack, D. (2018) *This PSA About Fake News From Barack Obama Is Not What It Appears* [article] Buzzfeed News, 17 April 2018.

7 See for example DensePose http://densepose.org/

2. How are deepfakes created?

Visual manipulation is nothing new. Photos and videos have been doctored since their very beginning. One of the earliest examples of face swapping can be found in a photo of Abraham Lincoln and US Southern politician John Calhoun, whose heads were carefully rearranged in a satirical gesture at the height of the American Civil War. But what has changed in recent years is the arrival of new forms of advanced software, including machine learning algorithms, which have made it easier and quicker to produce deepfakes. At the same time, this technology has been commercialised and mainstreamed through software like FakeApp and Face Swap.

With regards to **face replacement**, the developer Alan Zucconi sets out three steps in their formation: extraction, training and creation.[8]

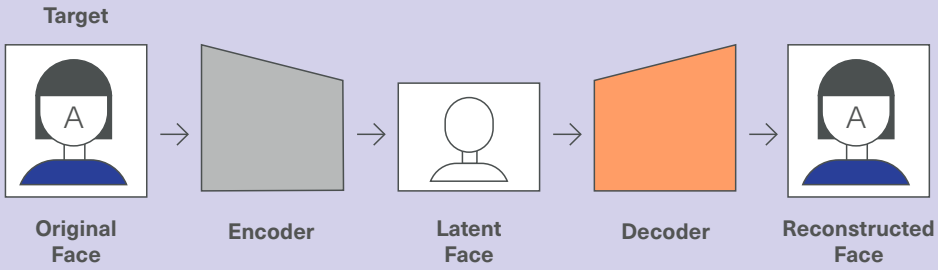
— **Extraction** – The first task is to collect sufficient images on which to train the face replacement model. A popular method is to start with videos of both the source and target persons, cut these into individual frames, and then crop the images so only a portrait of the face remains. Ideally the two individuals would resemble each other in face structure and head size. Software is now available to aid this extraction process.

— **Training** – The next task is to train the face replacement model using the images collected. This is done using an **autoencoder**, which is a neural network made up of two parts: an encoder and a decoder.[9] The encoder takes an image of a face and compresses it into a low dimension representation, also known as the 'latent face'. The decoder then takes that representation and reconstructs the face into its original form (see Figure 1).

8 Zucconi, A. (2018) *An introduction to DeepFakes* [blog] AZ website, 14 March 2018.

9 The models used in DeepFake software are known as Variational Autoencoders (VAEs), which is one variant of an autoencoder.

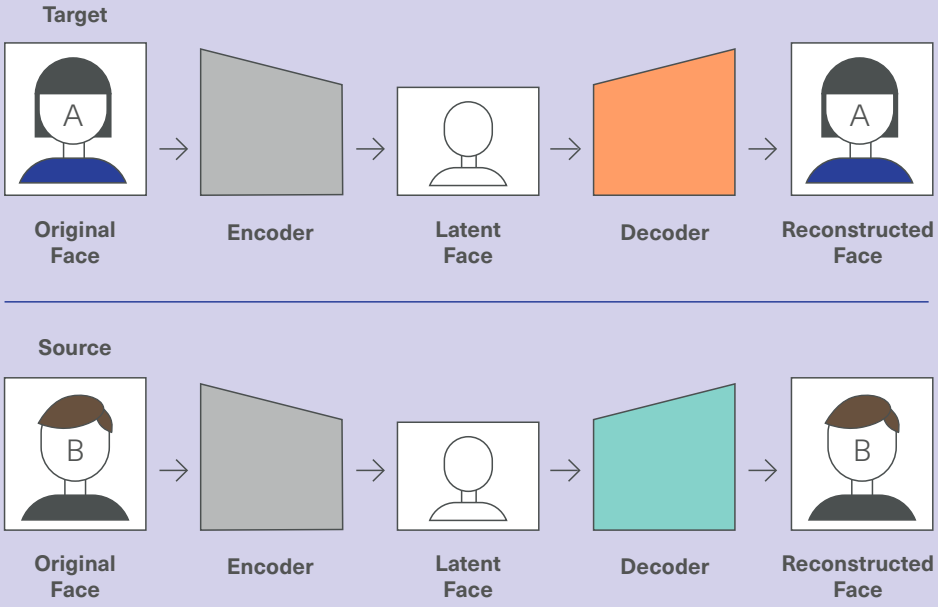
Figure 1



Source: Alan Zucconi

— Face replacement is made possible by training two of these neural networks – one for the source face and one for the target face. Both networks share the same encoder, which means that the decompressed version of both faces shares a similar baseline architecture. However, they have different decoders (see Figure 2). Both neural networks are left to train until the autoencoding process can reconstruct an image of a face similar to its original version.

Figure 2

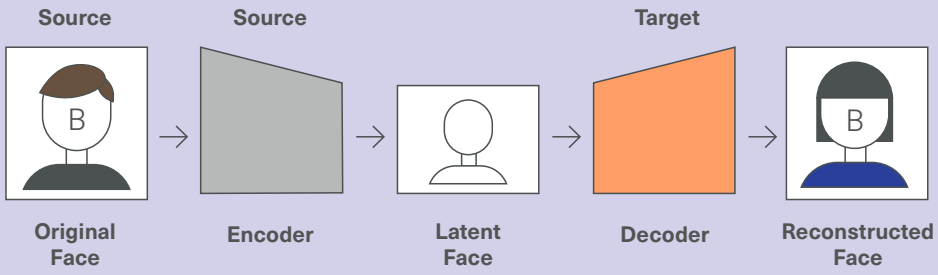


Source: Alan Zucconi

— Once the training is complete, **the decoders are then swapped**, so that the source image is decompressed using its original encoder but reconstructed using the target image's decoder. The result is an image that delicately stitches the source's face onto the target's, while staying true to the target's expressions (see Figure 3). Thus, when the target opens their mouth or moves their eyebrows, they do so but with the visual features of the source person.

— **Creation** – The final task, and the most technically challenging, is to insert these deepfake images into the desired video. This means ensuring that, for every frame in the video, the angle of the synthesised face matches the angle of the head of the target person. According to Zucconi, this is the only stage in the process that draws on hand-written code rather than machine learning algorithms, and is thus susceptible to more errors.

Figure 3



Source: Alan Zucconi

Face re-enactment, the other common form of deepfake, also draws on a type of auto encoder. However, this method requires only a single video of the target (e.g. a video of a news reader) rather than multiple individual images of both target and source, as is the case with the face replacement technique.

3. Are deepfakes a threat?

Deepfakes are viewed by many as a critical threat to individuals and society. *'You thought fake news was bad? Deepfakes are where truth goes to die'* reported *The Guardian* in November 2018.[10] *''Deepfake' videos threaten the world order'* ran a headline in *The Times* in February 2019.[11] Whether or not these fears are justified will depend on the quality of deepfakes, how they are used in practice, and the ability of the public to tell fact from fiction.

Face replacement in pornography has already caused distress to people who have been featured without their consent. Celebrities appear to have been the primary victims, but they are not the only ones affected. Indian journalist Rana Ayyub was left traumatised after her image was falsely implanted in a pornographic video that was shared more than 40,000 times.[12] Asher Flynn, Associate Professor of Criminology at Monash University, says these videos amount to **'image-based abuse'**, while academics at the Universities of Kent and Durham have argued that threats of this nature can be 'paralysing' for those involved.[13]

Others fear deepfakes will undermine our political discourse. Danielle Citron and Robert Chesney, two US law professors, imagine several scenarios where manipulated footage could destabilise democracies.[14] A deepfake showing a politician engaged in criminal activity may be enough to sway an election if released close to polling day, while a deepfake that falsely portrays a US soldier burning the Koran could result in worldwide protests. In May 2018, a deepfake was released in Belgium showing Donald Trump urging the country to pull out of the Paris Climate Agreement. Although it was a spoof film created by the Belgian Sp.a Party, the video was enough to rattle some viewers, forcing the Party to clarify that the video was indeed a fake.[15]

10 Schwartz, O. (2018) *You thought fake news was bad? deepfakes are where truth goes to die* [article] The Guardian, 12 November 2018.

11 Schick, N. (2019) *'deepfake' videos threaten the world order* [article] The Times, 27 February 2019.

12 India Today Web Desk (2018) *I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot* [article] India Today, 21 November 2018.

13 McGlynn, C., Rackley, E. and Johnson, K. (2019) *Shattering lives and myths: A report on image-based sexual abuse*.

14 Chesney, R. and Citron, D. (2019) *Deepfakes and the new disinformation war*. Foreign Affairs, January/February 2019 Issue.

15 Schwartz, O. (2018) Op cit.

As well as being disruptive in their own right, deepfakes could give bad actors **an excuse to deny involvement in untoward activity.** If every video has the potential to be a fake, it offers people the opportunity to challenge the veracity of genuine footage. This in turn could have consequences for how evidence is used in criminal investigations. Visual communications expert Paul Lester believes that "images – whether still or moving – will soon not be allowed in trials as physical evidence because of the threat to their veracity by digital alterations."[16] If deepfakes take off, it could force us to prove not just what is false but also what is true.

As with every technology, different groups face different threats. A number of our interviewees argued that Internet users in developing countries, where digital literacy remains relatively low, could be more susceptible to believing deepfakes. It has been claimed that the violence inflicted on the Rohingya community in Myanmar was partly fuelled by falsified text-based posts that promulgated on Facebook.[17] Deepfakes could similarly be used as a tool of subjugation by totalitarian regimes, or as a way for extremist groups to stir up social division.

Not everyone agrees with these claims, however. Sceptics argue that the fears surrounding deep fakes mostly relate to what *might* come to pass rather than what already has. The reality is that it remains difficult to create convincing forgeries, even for seasoned users of the underlying technology. Developer Alan Zucconi says that assistive software packages like FakeApp are straightforward to use but tend to churn out low-grade footage.[18] Much has been written recently of the Chinese app Zao, which allows users to easily plant an image of their face onto characters within famous movie scenes. However, the overall output is still crude, with the app only working for pre-modelled scenes rather than new ones created by the user. For now, the job of putting together a convincing deepfake involves multiple steps, and it only takes one slip up for the process to go awry. This may explain why we have yet to see a deepfake of a politician go viral.

Yet **deepfakes do not have to be flawless to inflict harm.** Many of the 8,000 deepfake pornographic videos that Dutch company Deeptrace says are on adult websites are likely to be of rudimentary quality, with most viewers knowing that the footage has been doctored.[19] However, this does not diminish the distress felt by those featured in the clips. One victim told Huffington Post that "the idea of people sexualising me makes me feel like I'm being fetishized, receiving unwanted attention, losing respect as

16 Gardiner, N. (2019) *Facial re-enactment, speech synthesis and the rise of the deepfake*. Retrieved from https://ro.ecu.edu.au/ theses_hons/1530

17 Mozur, P. (2018) *A genocide incited on Facebook, with posts from Myanmar's military* [article] The New York Times, 15 October 2018.

18 In CDEI interview with Alan Zucconi, undertaken on 25 March 2019.

19 Deeptrace (2018) *The State of Deepfakes: Reality under attack*.

a person and no longer safe."[20] The same feelings will be shared by those women targeted by the DeepNude app (now discontinued), which created crude images 'imagining' what people looked like undressed.

Some of the most effective forms of media manipulation rely on very little, if any, artificial intelligence. **'Shallowfakes'** are videos that have been edited using only basic techniques, such as changing the caption above clips or altering the speed of footage. US Senator Nancy Pelosi was recently targeted in a shallowfake, when footage of her speaking at a panel event was slowed down to give the impression she was inebriated and slurring her words.[21] A similar attack was made on CNN reporter Jim Acosta. In his case, footage of him asking a question at a Presidential press conference was sped up to falsely portray him as being aggressive to a White House staffer. Neither of these edits required specialist skills, yet the videos received widespread coverage nonetheless.

It is possible, of course, that growing media literacy among the public will subdue the impact of both deepfakes and shallowfakes. When Adobe introduced Photoshop in 1990, many thought it would be used to oppress the truth and hide misdeeds.[22] Yet for the most part, viewers learned to live with this technology by recalibrating their view of photos, shifting away from an assumption that 'seeing is believing'. While the same may happen with visual and audio content manipulated by AI and basic software packages, this cannot be assumed. Insufficient research has been undertaken to explore how people engage with doctored material and the effect it has on their beliefs and behaviour.

Box 1: Deepfakes for good

Deepfakes are not always deployed for malicious purposes. The film industry has long used visual and audio manipulation tools in post-production, and is now using the latest techniques to bring deceased actors 'back to life' for film sequels (a process known as 'digital resurrection'). The benefits of deepfake methods also extend into healthcare. Lyrebird, a maker of speech synthesis software, has partnered with a motor neurone disease charity to create digital copies of patients' voices, with the goal of creating artificial voice replacements that people can adopt once they lose the ability to speak.[23]

20 Cook, J. (2019) *Here's what it's like to see yourself in a deepfake porn video* [article] HuffPost US, 23 June 2019.

21 CBS News (2019) *Doctored Nancy Pelosi video highlights threat of 'deepfake' tech*.

22 Newsweek (1990) *When Photographs Lie* [article] Newsweek, 29 July 1990.

23 For more information visit https://lyrebird.ai/work

4. What should we do about deepfakes?

There are, then, several reasons to be wary about the use of AI and other forms of advanced software in doctoring audio and visual content. While sophisticated deepfakes remain difficult to produce, even rudimentary versions can cause harm. If the underlying technology continues to develop, as many expect it will, the barriers to producing deepfakes will fall and their quality will improve. In the last year alone we have seen the emergence of several new techniques for fabricating media content, particularly in the field of speech synthesis, where new models like RealTalk are demonstrating an impressive capability to generate audio clips of human targets.[24] Even if technological progress stalls, bad actors could still turn to shallowfakes in their attempts to distort reality on the internet.

We therefore need a robust system of governance – including appropriate regulations and screening tools – to ensure fabricated content is identified, contained and removed before it can inflict damage. Efforts at containment could take several forms:

Legislation

Legislators around the world are considering new laws to suppress audio and visual disinformation. In the state of New York, lawmakers have debated a new bill that would prohibit certain uses of a 'digital replica' of a person, on the basis that 'a living or deceased individual's persona is personal property'.[25] A similar bill is being discussed in the US Congress, which would create sanctions, such as jail time, for anyone found to be producing harmful deepfakes.[26] In 2017, the German government passed a new law introducing fines on tech companies that failed to take down racist or threatening content within 24 hours of it being reported – a law that could in time be targeted at deepfakes.

Critics, however, say using legislation to contain deepfakes would be both ineffective and counterproductive: ineffective because it is difficult to identify the makers of deepfakes, with many residing on foreign soil; and counterproductive because new legislation could have the unintended

24 Dessa (2019) RealTalk: *This Speech Synthesis Model Our Engineers Built Recreates a Human Voice Perfectly* [article] 15 May 2019.

25 Villasenor, J. (2019) *Artificial intelligence, deepfakes and the uncertain future of truth* [article] Brookings Institute, 14 February 2019.

26 Hao, K. (2019) *Deepfakes have got Congress panicking. This is what it needs to do* [article] MIT Technology Review, 12 June 2019.

consequence of curbing the use of visual and audio manipulation techniques for socially beneficial uses (see Box 1).[27] Legislation could also be used to stifle free speech under the auspices of tackling disinformation. Singapore recently passed legislation that will allow the government to order social media platforms to take down content that it considers to be false – a move that Human Rights Watch called a 'disaster for online expression'.[28]

Yet there is a role for legislation to encourage social media platforms to pay closer attention to the content being posted on their platforms. The new Online Harms White Paper sets out a credible and balanced vision in this regard, with a proposal to establish a new duty of care that would hold tech companies to account for addressing a range of online harms, one of which could be the spread of audio and visual disinformation.[29] There is also value in updating existing legislation to factor in deepfakes. The Ministry of Justice and the Department for Digital, Culture, Media and Sport recently commissioned a review of the laws around the non-consensual sharing of sexual images to ensure they capture new types of synthesised content created by advanced software, including deepfake pornography.[30]

Detection

Detection is another important containment tool. Media forensic methods have long been used in criminal courts to interrogate visual evidence, but they can also be applied to help identify deepfakes. One form of media forensics involves examining individuals in footage for physiological inconsistencies that arise from the way doctored videos are constructed. This includes looking at whether subjects blink during footage, and whether the colour and shadows on their skin appear to flicker. Another approach is to check whether the acoustics of a video correlate with the scene being recorded, for example the size of the room or the presence of people in the background.

Experts continue to disagree on whether media forensics is capable of screening out deepfakes. Some, like specialist Hany Farid, worry that bad actors will simply adjust their tools every time a new detection method goes public.[31] Others disagree and claim that detection methods are improving rapidly, particularly those that draw on AI. One AI-based tool called FaceForensics was trained on a dataset of half a million manipulated images, leading to detection results that are an order of magnitude better than human forensic teams.[32] Social media platforms have also implemented new screening procedures, including Facebook,

27 Chesney, R. and Citron, D. (2019) Op cit.

28 The Guardian (2019) *Singapore fake news law a 'disaster' for freedom of speech, says rights group.*

29 The Online Harms White Paper is accessible here: www.gov.uk/government/consultations/online-harms-white-paper

30 McGlynn, C., Rackley, E. and Johnson, K. (2019) Op cit.

31 In CDEI interview with Hany Farid, undertaken on 26 March 2019.

32 Rössler, A. et al. (2019) *FaceForensics++: Learning to Detect Manipulated Face Images*. Available here: https://arxiv.org/abs/1901.08971

which in September 2018 extended its proprietary fact-checking procedure to cover photos and videos alongside text.[33] Siwei Lyu at Albany University is testing a new method to prevent the re-use of online content as training data for deepfakes. This would involve inserting imperceptible 'adversarial noise' into images and videos, which could disrupt the process of face detection required to create deepfakes.

Also noteworthy are developments in provenance checking. Rather than interrogate the content of videos, these methods instead aim to verify their source, including where they were filmed and by whom. US start-up Truepic is one of several companies now selling software that adds a 'digital fingerprint' to images and videos at the point of their creation.[34] This information, which includes GPS data, is encrypted and stored as a file on Truepic's servers. A verification webpage is then created for the file, allowing forgery inspectors to view the original version and verify that it matches what they see in front of them. While watermarking of this kind shows promise, its effectiveness – as Danielle Citron and Robert Chesney point out – may depend on ubiquitous deployment across content capture devices, including smartphones and laptops.[35] It could also damage the anonymity of content creators, which includes minority groups facing suppression abroad.

Even were deepfakes (and shallowfakes) to be detected with complete accuracy, it is an open question as to whether they should always be removed from social media platforms. Facebook recently refused to take down a number of doctored video clips, including of Nancy Pelosi and Mark Zuckerberg, which it said did not violate its moderation rules. This points to the challenge of distinguishing malicious attempts to spread disinformation from harmless satire. Still, even if their findings are not always acted on, detection tools should at least be accurate and accessible to those who need them. Journalists in particular could benefit from screening aids as they attempt to discern fact from fiction in the content they view. The Wall Street Journal has reportedly formed a 20-person strong Media Forensics Committee to advise its reporters on how to spot doctored video footage, and has invited academics to give talks on the latest innovations in deepfake screening.[36]

Education

A third means of managing deepfakes is through public education. Although many people will naturally become more attuned to the presence of doctored footage and begin to view online content more critically as a result, some groups may benefit from having the phenomenon of deepfakes brought to their attention.

33 Woodford, A. (2018) *Expanding Fact-Checking to Photos and Videos* [article] Facebook Newsroom, 13 September 2018.

34 For more information visit https://truepic.com/

35 Chesney, R. and Citron, D. (2019) Op cit.

36 Southern, L. (2019) *'A perfect storm': The Wall Street Journal has 21 people 'detecting' deepfakes* [article] Digiday, 1 July 2019.

Deepfake researcher Vincent Nozick has talked of the importance of 'zététique', a French term that describes the critical analysis of paranormal activity that attempts to balance scepticism with open mindedness.[37]

The mainstream media has already helped to raise awareness of deepfakes, as Buzzfeed did in an article last year when it outlined 5 tips for spotting falsified videos.[38] Tech companies have also sought to educate the public. Synthesia, a creator of synthesised video content for use in digital marketing, is planning to produce educational content to improve video literacy among young people. Elsewhere, Google have created teaching materials aimed at supporting young adults to identify fake news online, which could be extended to cover visual and audio disinformation.[39] However, it remains to be seen whether content of this kind will have a bearing on people's ability to single out deepfakes in day-to-day internet use. To be effective, education must focus on 'evergreen' detection methods that people can use well into the future, rather than promote techniques that could fast become outdated.

37 Orange (2019) *Deepfakes, falsification of reality* [article] Orange Hello Future, 11 March 2019.

38 Mack, D. (2018) Op cit.

39 See Google's Be Internet Citizens programme: https://internetcitizens.withyoutube.com/

5. What can we do now?

Whether it is establishing new regulations, enhancing screening methods or raising public awareness, there are many ways to strengthen the oversight of deepfakes and shallowfakes. However, there is still much that is unknown about the effectiveness of different interventions, including their potential for unintended consequences. The government, tech companies and media forensics specialists must continue exploring and piloting new containment measures, while being mindful not to squeeze out beneficial uses of audio and visual manipulation. They will also need to do so at a pace that matches the speed at which the underlying technology progresses.

It is not for this briefing paper to make formal recommendations, however we would advise stakeholders to take the following steps as they attempt to create an effective governance regime for doctored content:

The UK government – The Department for Digital, Culture, Media and Sport (DCMS) may wish to include deepfakes in its on-going analysis of disinformation, so that visual and audio manipulation is captured in its analysis of other doctored content on the Internet. DCMS could also start and maintain a dialogue with deepfake experts such as those referenced at the bottom of this paper.

The research community – Research Councils could support new studies investigating the consequences of deepfakes for the UK population, as well as fund research into new detection methods. Research into visual and audio disinformation is sparse in comparison with that looking at written forms of disinformation. There is limited knowledge, for example, about which demographic groups are more susceptible to the messages conveyed by deepfakes.

Media outlets – Media outlets may wish to invest in deepfake detection tools and to boost their media forensics teams to ensure they are not inadvertently disseminating disinformation. At the same time, when reporting on deepfakes, journalists should be mindful of presenting a balanced account of their impact on society, remembering to comment on the beneficial uses of visual and audio manipulation alongside its potential to cause harm.

Technology companies – Technology companies, particularly those running social media platforms, could include audio and visual manipulation within their anti-disinformation strategies. They could also exchange best practice methods and tools for spotting deepfakes, including training data of manipulated sounds and images. In a welcome intervention, Facebook recently announced its intention to create and openly share a dataset of synthesised deepfakes, which will aid experts as they seek to hone their deepfake detection tools.

The general public – The general public should exercise caution when viewing audio and visual content where the trustworthiness of the source is in doubt. Social media platforms could offer guidance on how to sense-check videos and images, for example by checking the reliability of the source and looking for different versions of the content elsewhere on the internet.

This paper comments only on the near term risks of deepfakes. We therefore advise that researchers, journalists and policymakers continue to scan the horizon for technological breakthroughs and other developments that might change the overall threat level of audio and visual disinformation. Among them are:

— New production techniques that demand less training data or less powerful hardware

— The development of cheap, consumer-facing software packages that allow individual actors to create sophisticated content

— Breakthroughs in speech synthesis, which can be coupled with face replacement or face re-enactment techniques to create highly convincing videos

— The volume and quality of academic papers on deepfake detection methods, and the amount of funding coming from research councils, foundations and other donors

— The resources committed by social media platforms to improving deepfake detection

Table 1. Common myths surrounding deep fakes

	Myth	Reality
#1	Deep fakes predominantly take the form of face swapping in videos.	There are four main types of deep fake: face replacement (aka face swapping), face re-enactment (aka puppetry), face generation, and audio synthesis.
#2	Deepfakes are cropping up in large numbers on social media platforms.	Few political deepfakes have emerged on social media platforms. However, the number of deepfake pornographic videos is a cause for concern.
#3	Anyone can make sophisticated deepfakes that pass the bar of believability.	While supportive software like FakeApp has allowed more people to try their hand at deepfakes, high quality audio and visual synthesis still requires considerable expertise.
#4	The best way of detecting deepfakes is through physiological testing (e.g. the 'eye blinking' test).	Physiological examination of videos can be slow and unreliable. Systematic screening of deepfakes will require AI-based tools that can partially automate the detection of forged content. Tools will also need to be regularly updated.
#5	New legislation is a quick solution for dealing with deepfakes.	Attempts to legislate against deepfakes may prove ineffective, given it is very difficult to identify where doctored content originates. Legislation could also threaten beneficial uses of visual and audio manipulation.
#6	Deepfakes are just like photoshopped images. People will get used to them.	This is an assumption, not a fact. There has been insufficient research on how deepfakes influence the behaviour and beliefs of viewers.

About the CDEI

The adoption of data-driven technology affects every aspect of our society and its use is creating opportunities as well as new ethical challenges. The Centre for Data Ethics and Innovation (CDEI) is an independent advisory body, led by a board of experts, set up and tasked by the UK Government to investigate and advise on how we maximise the benefits of these technologies.

The CDEI has a unique mandate to make recommendations to the Government on these issues, drawing on expertise and perspectives from across society, as well as to provide advice for regulators and industry, that supports responsible innovation and helps build a strong, trustworthy system of governance. The Government is required to consider and respond publicly to these recommendations.

We convene and build on the UK's vast expertise in governing complex technology, innovation-friendly regulation and our global strength in research and academia. We aim to give the public a voice in how new technologies are governed, promoting the trust that's crucial for the UK to enjoy the full benefits of data-driven technology.

The CDEI analyses and anticipates the opportunities and risks posed by data-driven technology and puts forward practical and evidence-based advice to address them. We do this by taking a broad view of the landscape while also completing policy reviews of particular topics.

More information about the CDEI can be found at www.gov.uk/cdei and you can follow us on twitter @CDEIUK

Centre for
Data Ethics
and Innovation