How great is the threat?

Cyber EPQ

Martina King

GWC5530

Alleyn's School

17/03/2023

Abstract

Deepfakes have the potential to turn our society upside down. From embarrassing individuals, to scamming corporations, to overthrowing governments, advancements in technology make the potential threats almost limitless. But all hope is not lost. While there is likely to be a cat-and-mouse game between ever deeper fakes and ever more sophisticated detection methods, the real defence against the erosion of confidence in the reliability of media lies elsewhere. Blockchain-based record-keeping may protect individuals against fraud. Legislation can help protect society by limiting the creation and distribution of malicious synthetic media. And above all, a greater general awareness of the risks of fabrication, and vigilance when online, will lead the charge in the war against misinformation.

Contents

Introduction

From education to medicine, phishing scams to political tactics, deepfakes are ushering in a new era in technology. My essay will begin by covering what deepfakes are, including how they are made, and how the technology that creates them is becoming more advanced. I will then detail some of the ways deepfakes can benefit us, before I elaborate on the specific ways they may cause harm. Whilst a deepfake can have severe consequences for an individual, in my opinion we should be more worried of the risks that deepfakes pose to societies as a whole. The greatest is the likely erosion of trust in any form of media, and the potential loss of confidence in what constitutes fact or even truth itself. Video and audio recordings will no longer be valid as a form of testimony; anyone accused because of a piece of media can simply cry "Deepfake!" and that media becomes no more credible than any other human witness. How well we can reliably and accurately detect deepfakes will then be discussed, and then I will suggest different methods we can use to protect ourselves from deepfakes. I will argue that while the threat from deepfakes is very significant, it may yet be countered with a combination of legislation, detection and – most importantly – education.

Deepfake technology

Manipulation of media has been around since the 1860s [1]. The most notable early example is Stalin's intensive photomanipulation program, where thousands of photos were painstakingly edited for Stalin's personal political benefit [2]. But even earlier examples exist: the Ancient Romans were known to chisel names off stone, erasing any trace of certain individuals' existence. These early examples constitute cheapfakes – or shallowfakes as they are also known. These are similar to deepfakes, but are usually created by editing previous content as opposed to completely creating new synthetic media.

Deepfakes, in contrast, are a new form of media manipulation. They aren't just the new version of Photoshop: Photoshop requires a lot of human time and effort in the creation of an image. Deepfakes differ because they are created using artificial neural networks – a computer architecture loosely modelled on the human brain – that can teach itself how to achieve tasks.

An example of a neural network system that you'll probably be familiar with is facial recognition. When that neural network is trained, a dataset of faces and matching names are used as an input, then the system selects a face and tries to determine the name matched to it. The system then compares the result to the true one, if it is wrong figures out why it is wrong, then tries again with this new knowledge in mind. In this way the neural network can iteratively improve and learn by itself. This means that, compared to a normal algorithm that someone programs, the logic that neural networks follow doesn't make sense to humans. That is what makes them so powerful. They are a technology that surpasses human understanding.

Early image manipulation falls into one of four main categories (Figure 1). These are:

- Entire Face Synthesis, where completely new facial images are created;
- Identity Swap, where one person's face is overlaid on another's;
- Attribute Manipulation, where, for example, the colour of someone's hair could be changed;
- Expression Swap, where, for example, a happy face could be made sad [3].

There are also other less common examples: face morphing, face de-identification, and audio-to-video and text-to-video synthesis. All these types of manipulations are generally formed on a frame-by-frame basis, so whilst each frame individually may look realistic, the transitions between frames won't be seamless. This means they're easily detectable to the human eye.

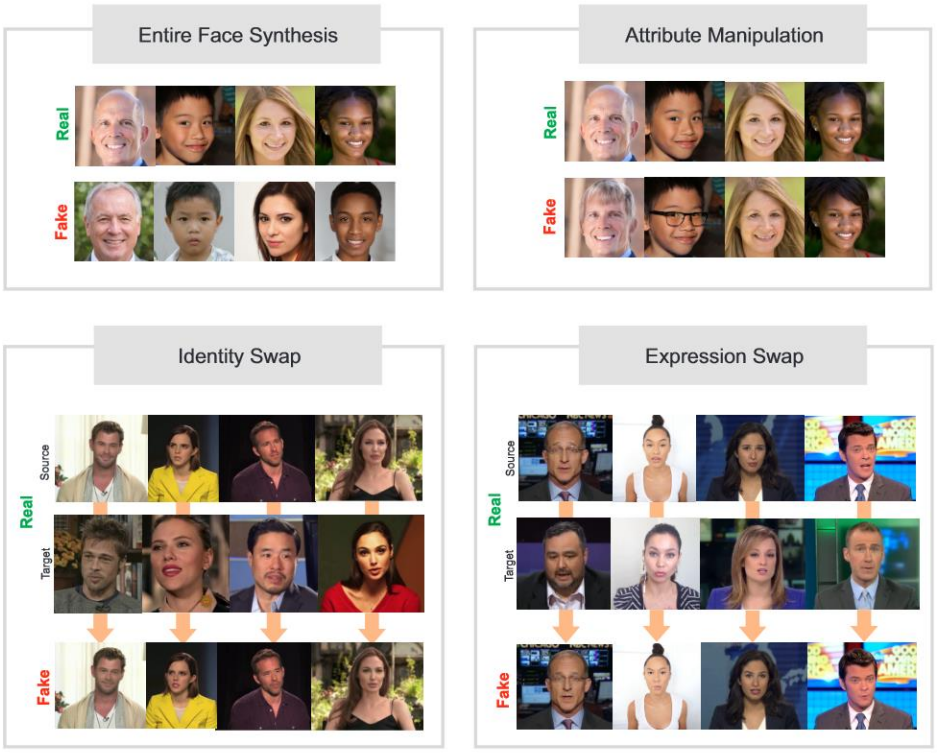


Fig.1. Examples of different types of image manipulation. Source: [3]

The more recent deepfakes are more realistic and much harder to detect. This is because they are created using generative adversarial networks (GANs), invented by Google researcher Ian Goodfellow [4]. They comprise two separate neural networks. The first is called the "generator" which creates a deepfake based on a dataset of videos and/or audios. The second, called the "discriminator" is shown the generator's deepfake, as well as an original one from the dataset. It then has to determine which one of the two media are fake. If it correctly identifies the fake, the generator can then use this information to create a new, harder-to-detect, deepfake. In an iterative fashion the generator and the discriminator work together until the deepfake becomes indistinguishable from the original dataset. The use of GANs permits faster and more accurate creation of deepfakes. By themselves, neural networks are only half of the equation. Without generative adversarial networks, deepfakes would not be as realistic as they are. [5]

GANs do have one limiting factor. They can only be as diverse as the training dataset that they are given, i.e. if the original dataset only had images and videos of men, the GAN would not be able to create a realistic deepfake of a woman. However, as people are uploading more and more images and videos of themselves online, the datasets available to anyone wishing to create a deepfake become vastly populated. The greater the size of the training set, the more

powerful and accurate the deepfakes are. This is why some of the best or rather most realistic deepfakes are those of famous celebrities or politicians.

Deepfake technology is also improving rapidly, to the extent that you don't need great technological skill to create a realistic deepfake. Anyone with access to the internet can. In fact, (with permission) I managed to create a deepfake of my headteacher within a weekend – which includes searching for and deciding which different tools I wanted to use and testing different audio and source videos. Many of the deepfake creation systems that are currently available require only a source video and a target face/audio. Yet as the technology quickly improves you may not require even this: simply specific who you want to be doing what and where, and a realistic video could be provided for you almost instantly.

Now if I was able to create a reasonably realistic deepfake in about six hours, can you imagine what professionals can do in six months with teams of people working on it? I think we must assume that several extremely well-funded national intelligence services are also hard at work on their own in-house versions of deepfakes and detection tools too.

Positive uses of deepfakes

In order to evaluate the threat that deepfakes pose we must first explore what benefits they could have. There already exist numerous examples of deepfakes being used positively, from entertainment to education to medicine.

For example in *Rogue One: A Star Wars Story*, actor Peter Cushing appears in his role as Grand Moff Tarkin, despite having died over 20 years before the movie was filmed! [6] Disney managed to use CGI techniques to essentially bring the actor back to life to play his role. Not only has Disney managed to resurrect actors, but they can now remove the problem of actors only doing age-appropriate roles – also in *Rogue One*, Disney created a 19-year-old Carrie Fisher (who was 60 at the time). However using these CGI techniques was hugely expensive and time-consuming. But with deepfakes that process will become so much easier.

Deepfakes can also be used for better dubbing of films, to allow politicians to speak in different languages and dialects, and in the health sector to give those who are about the lose their voice the chance to digitally clone it for their own use later in life. Further details on these additional positive uses of deepfakes are in Appendix 1.1.

Deepfakes don't need to have some greater purpose – they can also be used for fun. Any face-swapping tools or many Snapchat filters use deepfake technology. But my personal favourite use of deepfake technology is *Deepfake Neighbour Wars,* a TV show where it looks as if we get to watch the disputes that occur when celebrities live side-by-side. The intent of the show is not to deceive anyone into realistically thinking that Harry Kane had his patio tile cracked by Stormzy [10], but simply for entertainment.

Even these positives come with associated drawbacks. This does set a dangerous precedent for actors in the future, and resulted in a lot of backlash for Disney [7]. Why should we waste time and money trying to raise new celebrities if we can resurrect and de-age existing famous ones? Furthermore, it raises the question of who owns the right to their face? Grand Moff Tarkin's appearance was even described as "digital indignity" by *The Guardian*. [8] Did Peter Cushing consent to being in Rogue One, despite the VFX supervisor's insistence that he wouldn't have objected? [9] How can we know? What if instead of having to hire a famous

actor to do multiple films, you only had to pay for their face which could then be overlaid onto a person with a similar body? This would eliminate physical constraints of filming – one actor could be in many films at the same time. The issue of stunt doubles having to look similar to their actor counterparts could also be solved – someone with a similar body type can do the stunt, then the actor's face can simply be deepfaked onto the stunt-double's – et voilà!

Deepfake technology can be used positively to benefit anyone and everyone in almost every aspect of life. Yet it follows that they can be used for malicious purposes as well. The reason why they pose such a threat is because of our inherent trust that videos are an accurate representation of the world and the fact that videos can still have an effect even if we know them to be false. See Appendix 1.2 for more detail.

Threats posed by deepfakes

Deepfakes present a whole host of different types of threats, both to individuals and to societies as a whole. But the most significant threat comes from the potential loss of video and audio as testimony and from the subsequent erosion of trust that will inevitably occur.

Deepfakes are limited only by our imagination. Figure 2 contains a list of some possible deepfakes that could be used to harm society created by Chesney and Citron [11], who believe "there will be no shortage of harmful exploitations".

In more general terms deepfakes can be used to manipulate elections, irreparably damage reputations, interfere in nation military and/or intelligence operations. From a social point of view deepfakes can reinforce and exacerbate underlying social divisions. Deepfakes also pose threats to psychological security and political stability. Pantserev explains how "mass disinformation and the production of fake news has become the key instrument of modern psychological warfare" [12].

Deep fakes are not just a threat to specific individuals or entities. They have the capacity to harm society in a variety of ways. Consider the following:

- Fake videos could feature public officials taking bribes, displaying racism, or engaging in adultery.
- Politicians and other government officials could appear in locations where they were not, saying or doing things that they did not.[107]
- Fake audio or video could involve damaging campaign material that claims to emanate from a political candidate when it does not.[108]
- Fake videos could place them in meetings with spies or criminals, launching public outrage, criminal investigations, or both.
- Soldiers could be shown murdering innocent civilians in a war zone, precipitating waves of violence and even strategic harms to a war effort.[109]
- A deep fake might falsely depict a white police officer shooting an unarmed black man while shouting racial epithets.
- A fake audio clip might "reveal" criminal behavior by a candidate on the eve of an election.
- Falsified video appearing to show a Muslim man at a local mosque celebrating the Islamic State could stoke distrust of, or even violence against, that community.
- A fake video might portray an Israeli official doing or saying something so inflammatory as to cause riots in neighboring countries, potentially disrupting diplomatic ties or sparking a wave of violence.
- False audio might convincingly depict U.S. officials privately "admitting" a plan to commit an outrage overseas, timed to disrupt an important diplomatic initiative.
- A fake video might depict emergency officials "announcing" an impending missile strike on Los Angeles or an emergent pandemic in New York City, provoking panic and worse.

Fig.2. List of examples of threats that deepfakes could pose to societies. Source: [11]

Deepfakes mainly target individuals, with the risks being directly proportional to the popularity and power that said individual holds. A deepfake could be created of a politician meeting with foreign spies, or conducting criminal behaviour, or any of a whole host of things that could be irreversibly damaging to them. In 2019 a shallowfake was created by slowing down a video of Nancy Pelosi, resulting in more slurred speech to make it seem as if she was drunk at the time [13]. Deepfakes can be used to fake announcements from people in power,

whether that's Putin announcing a nuclear strike or my headteacher cancelling school for the day.

But it isn't just famous people that can be targeted. Deepfakes are becoming increasingly easy to create so can be used by anyone against anyone. As Michael Grothaus puts it: "deepfake technology's power lies in the fact that anyone can wield it regardless of their computing comprehension."[14] The possibilities are endless. It could be a petty act of revenge by a friend, or a fraudulent kidnapping claim. Depending on the circumstances, timing and circulation of the fake, the effects could be devastating [11]. Deepfakes could be used to bypass iris and voice recognition systems, or even in phishing scams. Imagine if instead of a dodgy-looking email sent supposedly from a friend, you were actually hearing their voice tell you to click on a link or send them some money. In March 2019 a deepfaked voice was reportedly used to convince the chief of a UK subsidiary of a German energy firm to transfer €220,000 to a Hungarian bank. The UK chief recognised the "slight German accent and the melody" [15] and subsequently transferred the money. The caller attempted to trick the UK chief a couple more times into sending more money, however the chief had grown suspicious and did not make any more transfers [16].

Narcissistic abusers could also uses deepfakes to gaslight their victims. As explained by Rini and Cohen: "fake photographic and video evidence can be used to manipulate autobiographical memories" [17]. If someone is already mentally vulnerable, why should they trust their own impaired memory compared to a seemingly more reliable video or audio recording?

Furthermore even if people know that a video is false, it can still cause harm. The Guardian suggests that "deepfakes don't need to be undetectable or even convincing … to do damage" [18]. Shared by Trump himself, another shallowfake of Nancy Pelosi, this time ripping up his 2020 State of the Union speech (which she had done, but 50 minutes later than the shallowfake showed), led to huge uproar amongst his avid Twitter supporters [19]. Even though many knew the video to be fake, it still spread like wildfire over social media. The harms that deepfakes could pose are only exacerbated by the tendency of social media to spread false content much more than real content. In fact, in a study by 3 MIT scholars it was seen that "false news stories are 70% more likely to be retweeted than true stories." [20]

Despite the previously mentioned individual uses of deepfakes, I believe that the greatest threat that they pose is the reduction in our ability to rely on video and audio recordings as a form of testimony. Currently video and audio recordings are used to verify what people say. Let's say someone claims that they were working in a café at the time when a crime was committed, but CCTV footage shows that they weren't actually in the café. We would naturally all assume that they were lying. Or if CCTV footage showed that they were in the café, we would assume that they were telling the truth. In both cases we are assuming that what the video showed was the truth. But if suddenly videos can be believably faked, we can no longer rely on the video to represent the true events that occurred. The problem arises because we view videos as an "epistemic backstop" meaning that we trust that videos are an accurate representation of the real world [21]. Yet with deepfakes we lose this trust. Furthermore even if you can prove that a video is a deepfake you will never be able to prove that one isn't. Anyone who is convicted or accused of something only because of a video or audio recording could simply claim that it was a fake (a phenomenon Chesney and Citron call "the liars' dividend"), and the recording becomes no more reliable than if a witness had

claimed it themself. As Rini believes: "the gravest danger of deepfakes … [is] that they will gradually eliminate the epistemic credentials of all recordings" [21], to the extent that we can no longer trust anything that we see or hear online. Imagine what will happen to the justice system when video and audio recordings become no longer admissible in court.

Deepfake detection methods

Since deepfakes gained popularity and have become much more widespread, the demand for fast, accurate and reliable detection tools has skyrocketed. Many different types currently exist, however none provides the perfect solution. The demand for a perfect deepfake detection tool is so great that the United States government has even offered a reward of $5million to anyone who can provide a sufficiently fast, accurate and reliable tool. [14] Personally, I think that there will never exist one single tool that can concretely prove a video to be false, but that we will instead have to rely on lots of different tools to come to a balanced conclusion as to whether we think the video is fake.

Before deepfakes, manipulated content was detected by fingerprint analysis, analysing colour filters, interpolation, image compression and assessing possible editing of various image fragments and lowering the frame rate. However, traditional forensic approaches for digital video content turned out to be poorly acceptable since fakes were made by automatic tools. [22] Initially, most detection methods started off using convolutional neural networks (CNNs) – a type of neural network used to analyse visual imagery – which would generate predictions based on a single frame of a video. These predictions would be made using techniques such as facial recognition, which also uses neural networks. The tool tries to detect inconsistencies in each frame, then outputs a combined probability of whether the media is fake.

The problem with the detection algorithms described above is that they look at frames only individually and are therefore blind to any so-called "spatio-temporal" inconsistencies that occur. These inconsistencies are (mostly) visible to the human eye but not to neural networks. Subsequently, most deepfakes have a "lack of temporal coherence that is not noticeable in individual frames"[22]. As a result, there are now many new detection methods that analyse multiple frames across a period of time and try to identify inconsistencies. These spatio-temporal detection techniques look at local and/or global inconsistencies in the video. Local inconsistencies could include seemingly briefly having two sets of upper teeth or two lower lips, whereas global inconsistencies include finding portions of a video that have been unnaturally edited into a longer video. (Figure 3)

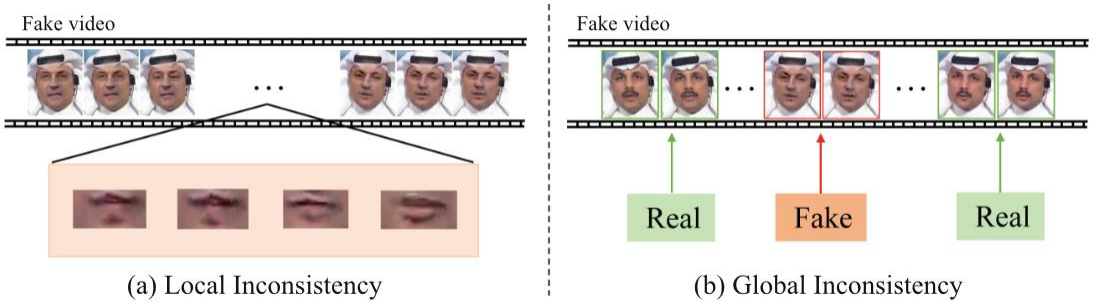


Fig.3. Illustration of local and global inconsistency. The former refers to the irregular facial movements within several consecutive frames. The latter stands for the cases of partial manipulation in the video. Source: [23]

An example of an inconsistency visible only over time is aberrations in eye-blinking (Figure 4). Li et al describe a method of firstly identifying the frequency of blinking of someone in a video and then comparing it with average blinking rates for whatever activity that person is engaged in. Their logic is based on the idea that the dataset that most NNs use to create deepfakes contain images that the target has posted online, most of which are probably with their eyes open. Hence the resulting video that the NN creates is likely to have far fewer frames with the target's eyes open than closed, resulting in a video with a blinking rate far too low. Conversely, if for some reason the target blinked far more than usual in the video, that would also be detected and highlighted.

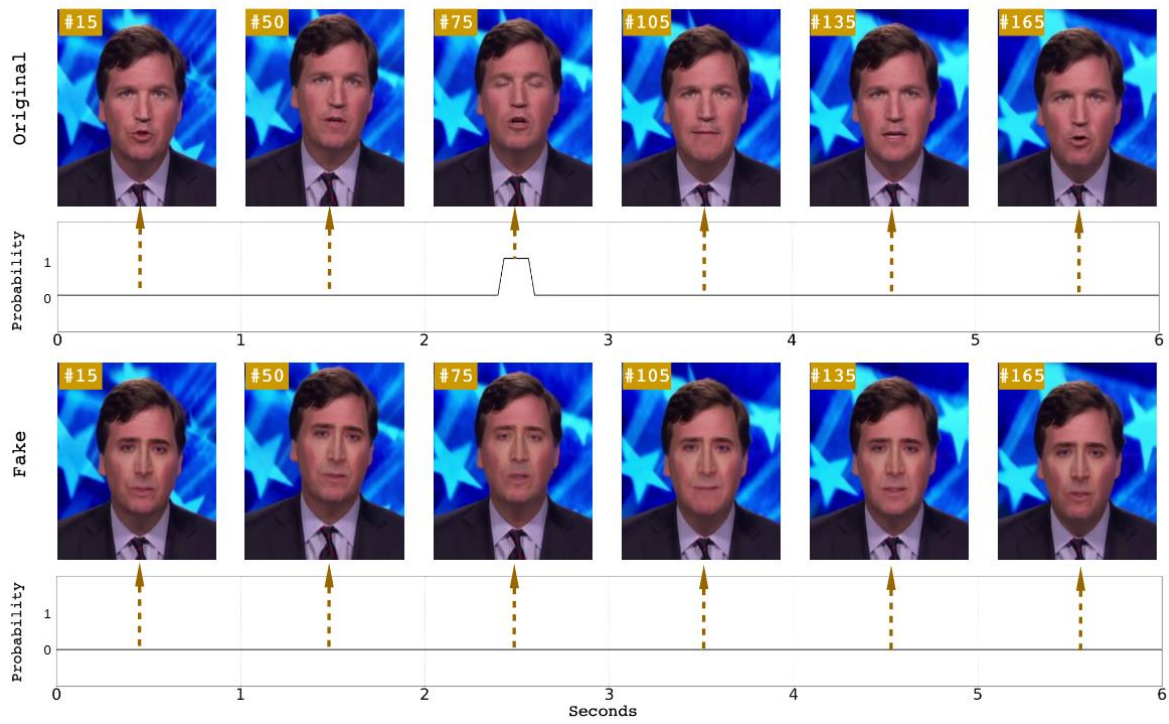


Fig.4. Example of eye blinking detection on an original video (top) and a DeepFake generated fake video (bottom). Note that in the former, an eye blinking can be detected within 6 seconds, while in the latter this is not the case, which is abnormal from the physiological point of view. Source: [24]

There exist many more different types of detection methods, such as detecting and recovering sequential deepfake manipulation [25]. This detects when deepfakes have been created by using multiple attribute manipulation techniques. Another example is detecting deepfakes by analysing image matching where a neural network tries to determine where multiple images have essentially been combined to produce the deepfake [26].

However, these detection methods are far from perfect. Given that most detection tools are trained on a specific dataset, if they encounter a deepfake made using a different method to those in the dataset they will inevitably fail. Each detection tool will recognise a deepfake only if it was made in a certain way. This suggests that in order to be able to detect a deepfake if we do not know how it was made, we must run it through many different detection tools before being able to pass a confident judgement on the source's veracity. For example, if the deepfake was created by lip-synching a video with new audio, trying to analyse

inconsistencies in eye-blinking wouldn't help at all. Moreover, many detection tools are only trained on deepfakes created without GANs, i.e. where the NN used to create the deepfake is not aware of detection methods. Hence these detection tools will often fail if presented with a deepfake created using GANs.

Naturally, there is the fear that deepfakes will evolve to be undetectable. In practice, however, as the technology to create them evolves and improves, so will the technology to detect them . I think that we will end up with an eternal cat and mouse game, with new deepfakes being used to train new detection technology. As one improves, so will the other. Pantserev states that "all of those methods that help in identifying a deepfake today will fail to discover it in the future" [12], and similarly, John Villasenor states: "deepfake detection techniques will never be perfect"[27].

I do not think that deepfake detection technology will be an instant and standalone solution to deepfakes' many threats. Whilst I believe that it is imperative that we develop reliable deepfake detection tools, I think that the best detection tool that currently exists is our own brains. Deepfakes have not yet become utterly indistinguishable from true videos, and until that point I think that we should be helping people learn how to identify deepfakes for themselves. Matt Groh and his colleagues at MIT created Detect Fakes – a website where you are presented with various pieces of media and have to say the extent to which you believe it to be real or fabricated – in an attempt to evaluate whether humans are more accurate at detecting deepfakes than machines. The more exposed you are to deepfakes, the better you become at detecting them. Groh lists possible things to look out for, from the texture of the skin to the lighting of the face, all in an attempt to help people become more vigilant and better versed in recognising deepfakes. [16]

Mitigating deepfake threats

At the moment there are 3 main possible routes to mitigating the threats posed by deepfakes. The first one is legislative i.e. governments create laws around the creation and distribution of deepfakes. The second would devolve responsibility to the private sector i.e. social media companies creating rules around the creation and distribution of deepfakes, and the third would rely on the individual themselves. I believe that this is the most important, as with greater personal vigilance can people better protect themselves from deepfakes.

In terms of legislation, California has introduced bill AB 602 that that makes it illegal to create or distribute videos, images, or audio of politicians doctored to resemble real footage within 60 days of an election [28]. Similar laws exist in Virginia and Texas, but no other US state. China was the first country to pass a national anti-deepfake law, making it a criminal offence to distribute deepfakes without disclosure, doing so on 1st January 2020 [29], and in January 2023 introduced a ban on non-consensual "deep synthesis technologies" [30]. The UK plans to make has sharing of non-consensual deepfakes showing explicit content illegal in the Online Safety Bill introduced in 2022 and currently in Parliament [31]. Whilst I believe that these are all steps in the right direction, I don't think that legislation by itself will resolve the issue of deepfakes. Not only is legislation notoriously slow in bringing about change, deepfakes can be argued to be an expression of free speech. Therefore any attempt to ban deepfakes would be seen to be violating this innate human right. As Caroline Kerner puts it: "there has been a recognised response to deepfakes in government, but these efforts are slow-

moving and limited in the scope of what they can achieve" [32]. Even if there were to be laws banning the creation and distribution of deepfakes, simply banning it won't solve the issue. Also discussed by Kerner is an open hearing on deepfakes and artificial intelligence that occurred in the US House Intelligence Committee in June 2019. This emphasised that the responsibility of labelling manipulated video should rest with the internet platforms who host the content.

I believe that companies who own and regulate platforms in which misinformation can easily be spread should have a responsibility to limit said spread. In addition social media platforms are more likely to have the technological skill to do so, as Kerner says: "it is likely that platforms have the expertise and the most direct ability to address video forgeries" [32]. Many such companies have already begun to do so. For example, through Twitter's introduction in 2020 of the "manipulated media" label. Many companies providing access to the general public with tools to create deepfakes have disclaimers saying their tools must not be used for malicious purposes. Yet I believe that, although helpful, this labelling will in no way hinder those who wish to use deepfakes for malicious purposes anyway. When attempting to clone my headteacher's voice (with her permission), in order to submit the training data for the voice to then be cloned the software required a voice authentication of the person whose voice you wish to clone reading out a statement detailing how they permit the software to be used to clone their voice. This is merely an example of barriers that companies can put in place to try to limit the malicious use of deepfake software.

However, I do not believe that the responsibility should solely lie on the companies, but partly on the individual consuming the content as well. Personally I think that the best way we can mitigate the threats posed is through a greater general awareness of deepfakes. If people are more cautious about the content that they consume then they are less likely to fall into the trap of blindly believing everything that comes their way. In addition, the more you are exposed to deepfakes the better you become at detecting them yourself. In creating Detect Fakes, Groh said that "the best way to inoculate people against deepfakes is exposure" [16]. I believe that people should not just aim to be more aware of deepfakes, but news in general. People who are more critical of content and more politically aware are also less likely to fall prey to deepfakes. As Appel and Prietzel discovered: "people with high analytic thinking and political interest were better at identifying a fake news article to be inaccurate" [33]. Furthermore, with this greater awareness comes the acknowledgement that any content that you upload online can be used to create a deepfake. As Halsey Burgund, a fellow in the MIT Open Documentary Lab, puts it: "It's a time to be more wary…One should think of everything one puts out on the internet freely as potential training data for somebody to do something with." [33]. So not only should we be more conscious of the content that we consume, but also that which we post ourselves.

Another possible way to mitigate the threats posed by deepfakes is through finding a method of disproving one through means such as authentication trails through wearable tech and blockchain-based record-keeping. For example, each time a real video is filmed, a way could be devised for metadata to be intrinsically written into the recording, detailing information such as when and where the video was made, or even something like what the weather was like. I believe that ideas such as these will become much more common in the future as people (especially ones in power) become increasingly concerned about deepfakes and try to disprove any that might arise opposing them.

Conclusion

Deepfakes represent a new era in technology, one that I believe poses a significant threat to all of our lives. Whilst deepfake technology can be used for good, in my opinion the benefits in no way outweigh the risks. I also think that the extent of the threat posed by deepfakes depends on the individual. The more famous you are, and the more power you have, the more you should be concerned. Despite their undeniable threats, I do not think that all hope is lost. I believe that deepfakes gain their power from being wildly publicised on social media and/or easily believed. With increasingly powerful and reliable detection tools and the implementation of anti-deepfake legislation, we can take meaningful steps to ensure that we reduce the spread of misinformation online. And yet these steps might not be enough. We have to stop the general public from blindly believing what they see and/or hear on the Internet. Personally, I think that the best way to protect ourselves from deepfakes is through a greater general awareness of deepfake technology. If people are better educated and can learn to think more critically about content that they consume online, they are far less likely to fall prey to fakes of any kind – whether shallow, cheap or deep.

Bibliography

1. Waters, M. (2017). *The Great Lengths Taken to Make Abraham Lincoln Look Good in Portraits.* Atlas Obscura. https://www.atlasobscura.com/articles/abraham-lincoln-photos-edited
2. Blakemore, E. (2018). *How Photos Became a Weapon in Stalin's Great Purge.* History Channel. https://history.com/news/josef-stalin-great-purge-photo-retouching
3. Tolosana, R. et al (2020). *Deepfakes and beyond: A survey of face manipulation and fake detection.* Information Fusion, 64, 131-148.
4. Goodfellow, I. et al (2014). *Generative Adversarial Networks*. Advances in neural information processing systems, pp. 2672–2680.
5. Dack, S. (2019). *Deep fakes, fake news, and what comes next.* University of Washington. https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next
6. Itzkoff, D. (2016). *How 'Rogue One' Brought Back Familiar Faces.* New York Times. https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html
7. Walsh, J. (2016). *Rogue One: the CGI resurrection of Peter Cushing is thrilling – but is it right?* The Guardian. https://www.theguardian.com/film/filmblog/2016/dec/16/rogue-one-star-wars-cgi-resurrection-peter-cushing
8. Shoard, C. (2016). *Peter Cushing is dead. Rogue One's resurrection is a digital indignity.* The Guardian. https://www.theguardian.com/commentisfree/2016/dec/21/peter-cushing-rogue-one-resurrection-cgi
9. Pulver, A. (2017). *Rogue One VFX head: 'We didn't do anything Peter Cushing would've objected to'.* The Guardian https://www.theguardian.com/film/2017/jan/16/rogue-one-vfx-jon-knoll-peter-cushing-ethics-of-digital-resurrections

10. Heritage, S. (2023). *Behind the scenes of TV's first deep fake comedy: 'None of it is illegal. Everything is silly.'* The Guardian. https://www.theguardian.com/tv-and-radio/2023/jan/09/deep-fake-neighbour-wars-interview-itvx-comedy
11. Citron, D. K. & Chesney, R. (2019). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.* 107 California Law Review 1753.
12. Pantserev, K. A. (2020). *The Malicious Use of AI-Based Technology as the New Threat to Psychological Security and Political Stability.* Springer Nature Switzerland.
13. Sadiq, M. (2019). *Real v fake: debunking the 'drunk' Nancy Pelosi footage – video.* The Guardian. https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video
14. Grothaus, M. (2021). *Trust no one – Inside the world of deepfakes.* Hodder Studio.
15. Stupp, C. (2019). *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case.* The Wall Street Journal https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
16. Somers, M. (2020). *Deepfakes, Explained.* MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained
17. Rini, R. & Cohen, L. (2022). *Deepfakes, Deep Harms.* Journal of Ethics and Social Philosophy 22 (2).
18. Schwartz, O. (2018). *You thought fake news was bad? Deep fakes are where truth goes to die.* The Guardian. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth
19. Lerman, R. (2020). *Nancy Pelosi's office hits out at Twitter and Facebook over edited video of her tearing up Trump's speech.* The Independent. https://www.independent.co.uk/news/world/americas/us-politics/nancy-pelosi-trump-video-facebook-state-union-speech-deepfake-edit-a9326721.html
20. Dizikes, P. (2018). *Study: On Twitter, false news travels faster than true stories.* MIT News. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308
21. Rini, R. (2020). *Deepfakes and the Epistemic Backstop.* Philosophers' Imprint 20 (24):1-16.
22. Sebyakin, A. et al. (2021). *Spatio-Temporal Deepfake Detection with Deep Neural Networks.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-71292-1_8
23. Gu, Z. & Yao, T. et al. (2022). *Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19775-8_35
24. Li, Y. et al. (2018). *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.* University at Albany, SUNY.
25. Shao, R. et al. (2022). *Detecting and Recovering Sequential DeepFake Manipulation.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19778-9_41
26. Dong, S. et al. (2022). *Explaining Deepfake Detection by Analysing Image Matching.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19781-9_2
27. Villasenor, J. (2019). *Artificial intelligence, deepfakes, and the uncertain future of truth.* Brookings. https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth
28. Paul, K. (2019). *California makes 'deepfake' videos illegal, but law may be hard to enforce.* The Guardian. https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce
29. Bhattacharya, A. (2023). *China goes a step further in regulating deepfakes.* Quartz. https://qz.com/china-new-rules-deepfakes-consent-disclosure-1849964709

30. Kharpal, A. (2022). *China is about to get tougher on deepfakes in an unprecedented way. Here's what the rules mean.* CNBC. https://www.cnbc.com/2022/12/23/china-is-bringing-in-first-of-its-kind-regulation-on-deepfakes.html
31. https://www.gov.uk/government/news/new-laws-to-better-protect-victims-from-abuse-of-intimate-images
32. Kerner, C. M. (2020). *Detecting Deepfakes: Philosophical Implications of a Threat.* Bachelor's thesis, Harvard College.
33. Appel, M. & Prietzel, F. (2022). *The detection of political deepfakes.* Journal of Computer-Meditated Communication Oxford Academic. https://academic.oup.com/jcmc/article/27/4/zmac008/6650406
34. Christopher, N. (2020). *We've Just Seen the First Use of Deepfakes in an Indian Election Campaign.* Vice. https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Appendix 1:

Appendix 1.1: More positive uses of deepfakes

On the same theme of entertainment, deepfakes could be used for betting dubbing of films. Grothaus advocates for the greater "cultural awakening" that would occur if people had access to films that didn't seem dubbed [14]. Instead of dubbed films as we know them, deepfakes could make it to seem that those actors were actually speaking a different language. Lip-synching deepfake software would allow for many more people to broaden their minds to watch a whole host of films that they previously had dismissed due to the cognitive dissonance of watching a dubbed film. Deepfakes have already been used to allow someone to speak another language. In February 2020, Manoj Tiwari, President of the New Delhi state BJP faked himself speaking a dialect native to a region he was wishing to gain more votes from [34]. Many voters there were moved by his seemingly flawless language and were therefore more sympathetic to his views. This was the first example of deepfake technology being used in elections.

The games industry will also hugely benefit from deepfakes – which is the reason they are so invested in their development. Deepfakes can allow for much more realistic avatars in games that can be quickly and accurately customisable however the user wished. The film Ready Player One shows a good example of how we would be able to create these avatars. Instead of paying vast amounts of money to use a celebrity's voice in a game, a games company would only have to pay for, let's say, a 30min recording session to properly clone their voice, and then anything else that they want their character with the celebrity's voice to say, they can deepfake. Deepfake technology could also be applied to give virtual assistants corporeal form. Imagine waking up on a Sunday morning to see your mother on your phone telling you to clean your room – terrifying! Another example is online shopping. Deepfakes could mean no more returning of clothes that don't fit you as they fit the model on the website. With a tap of a button you could live-deepfake the clothes onto yourself. And that wouldn't just be beneficial to users but to the companies too as they save money on returning items and refunding people, as well as maintaining happier customers. With deepfake technology also arrives virtual influencers, who, like virtual actors' faces, would save vast amounts of money through bypassing all human physical constraints such as only being in one place at one time.

Deepfakes are already helping people struggling with various illnesses. The ALS association are providing for people who have early ALS to clone their voices so that when they eventually lose the ability to speak, rather than having a robotic text-to-speech character say their words, they can quite literally say it themselves. Education can also be improved through deepfakes, for example "it will be possible to manufacture videos of historical figures speaking directly to students, giving an otherwise unappealing lecture a new lease on life" [11] – "a scene from a war film could be altered to make it seem that a commander and their legal advisor are discussing the application of the laws of war".

Appendix 1.2: Why deepfakes pose such a threat

I believe that the mean reason why deepfakes pose a threat in the first place can be boiled down to the phrase "seeing is believing". That is to say that deepfakes are so dangerous is because when we see a video or hear an audio recording, unless we see or hear some obvious editing we inherently believe it to be true. Rini explains how "video and audio recordings function as an epistemic backstop… Our awareness of the possibility of being recorded provides a quasi-independent check on reckless testifying thereby strengthening the reasonability of relying upon the words of others"[21]. She goes on to say that "recordings do this in two distinctive ways: actively correcting errors in past testimony and passively regulating ongoing testimonial practices" and Kerner agrees, saying "audio-visual recording is widely considered the ultimate authoritative evidence. It serves as an authority for real-world events and a tool for accountability," [32]. With deepfakes arrives the loss of this function of media such as audio and video. We will no longer be able to use media as reliable sources of true accounts of events unless we are confident in their validity which I will discuss the impacts of later.

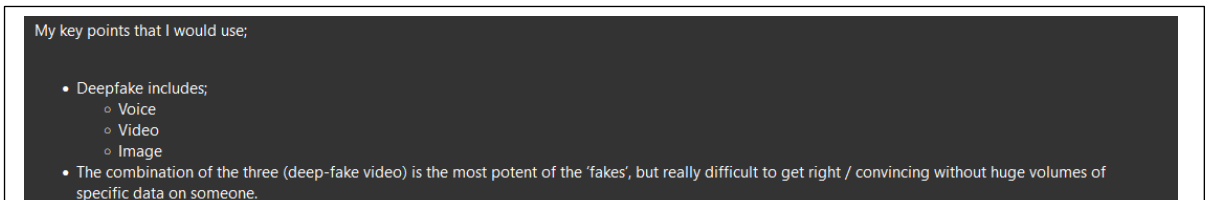
But first I wanted to explain an often overlooked aspect of deepfakes, which is their phenomenological impact. Phenomenology is the philosophy of experience and I believe that the threat that arises from deepfakes can be summarised by Maya Angelou's famous saying of "people will forget what you said, people will forget what you did, but people will never forget how you made them feel." That is to say that even if someone is aware that a video is a deepfake, the contents will still have had an impact on them. Rini states that "a deepfake video needn't be positively believed by viewers in order to be effective" and Kerner explains how " due to the phenomenology of video, human engagement with the medium is not purely epistemic." What I believe Kerner is saying is that if you were to tell a machine that something isn't true, it would simply discard that information and think no more of it. However we are humans and not machines, so cannot simply erase the existence of a deepfake from our minds on command.

Appendix 2:

Overview of Interview:

Whilst beginning my EPQ I was worried as to whether I would be able to find an interviewee, however within 24 hours of emailing Nicky Bodily I managed to find and organise an in-person interview for the week with Pete Trainor.

The interview was incredibly useful as we discussed things like the use deepfakes in the games industry that I hadn't previously considered. See the email in Figure 5 for a summary of the topics discussed.


My key points that I would use;

- Deepfake includes;
  - Voice
  - Video
  - Image
- The combination of the three (deep-fake video) is the most potent of the 'fakes', but really difficult to get right / convincing without huge volumes of specific data on someone.

To make reviewing the interview easier I also recorded it. Below is the transcript.

Transcript of Interview:

P.T: Can you focus on, you are focusing on, the video output?

M.K: Well, both, but yeah.

P.T: But the, um, uh, replicating of lots of different things by machines is also considered deep fake.

M.K: Yeah.

P.T: So, um, you know, generating words on behalf of somebody else is considered deep fake, um, audio is probably going to be far more dangerous in the long run because of, you know, prank phone calls. Or sound bites generated by politicians or whatever. So we're on the ethics bit by that. But, um, yeah, try and I would, I would set your stall out for this kind of work as kind of, um, maybe you are gonna focus on the video aspect cause that's kind of more interesting. I've got some good examples, but actually deep fakes span quite a, a wide range of media and, um, and some of that stuff is, is, is pretty, um, I think more interesting in some ways. So one of the things that I did think would be interesting for you to think about was, um, where a lot of this technology comes from. Yeah. So a lot of this technology is actually originated from, um, like the games industry. Yeah. So if you think about like computer games as a, as a medium, a lot of money has been pumped into. This, this thing I'm showing you here is, um, this is a relatively new one, but it's the same technology that's been used for like, maybe 30 years, which is a Sprite. Um, that's been designed by designer, but it's actually reading copy that you can put in, so it will replicate whatever copy you put into it in real time.

M.K: That's so cool. And that's how they generate all those videos in computer games and things like that. Yeah.

P.T: So the quality is computer games quality, but actually what's really interesting, so this is this one, just turn the sound up is, uh, I literally just copy and pasted, uh, some really weird copy off a off a Bitcoin website. "Are you looking for a fun and exciting way to earn real Bitcoin? Look no further than Bitcoin miner. This addictive mobile game…" So that's just copy that's come from over there and that's basically a malleable avatar. That's kind of the thing that's used in computer games.

M.K: Yeah. Wow.

P.T: Right. So Japan is really the kind of, they, they start this kind of process years and years ago, and games graphic starts getting really good and they realize that what they can't afford to do is generate almost like animate, you know, seen by scene, everything lip synced to famous actors or whatever. So they create the kind of technology that will, you know, uh, a 3D model that you can then push text through and it will kind of lip sync to that. Those are really the early versions of what is now the kind of modern, deep fake technology, and it was a necessity for that. Now, the other thing that's really interesting about the games industry that's worth noting, uh, as part of your work is that they were also the first people to, um, replicate people's voices. So there was a big thing for a while of famous kind of Hollywood actors, um, VO doing voiceover for computer games, um, and a lot of, um, like episodic computer games as well. There's a lot of, um, kind of voiceover work in things like the last of us, I think, which is that TV thing that's on the moment. But the game of that, um, what they would do, what the games industry would do is get famous actors and actresses into the studio, have them record 30 minutes of dialogue, and then have their machines replicate their voice patterns for the anything else they wanted to do. That's just worth bearing mind as kind of the origins of some of this stuff. So it was all created with very pure intentions to kind of democratize gaming or create more, um, make it easier for games and things like that. Um, the other people that were like right early doors on some of this was Pixar. So Pixar obviously did a huge amount of, like, all their actors in Toy Story and stuff are, you know, Tom Hanks and, and blah, blah, blah. Um, They're getting those actors through their, their studios. They're doing voiceover for, you know, 30% of the movie, and then they're generating as much of that from voice patterns as possible and having their, their avatars, if you like, their computer generating sprites mimic and, and lip sync to that. Yeah. So the origins of this stuff have, are kind of really fascinating. It was never designed particularly to, um, you know, create some of the harms that have now been created using it. So what happens between, um, and another really good example, which you might wanna just reference early doors, um, is like the Rogue One. So the first big, um, ethical and kind of complicated one is, you know, Rogue One, you've got two actors that are dead. You know, you've got Carrie Fisher and Peter Cushing are no longer here. Yeah. If other actors play their parts and then they deep fake in, um, the technology. Now what what's really interesting here is that there's two versions on the, on the screen. Um, there's the original and then there's the deep fake. So the original is where they tried to do it, just computer animated. So they had somebody like do it sort of almost manually. And then there's the deep fake technology, which is much, much better, which is where basically an algorithm is doing all of that. So the quality difference between an actual animator doing it and a computer interpreting the content when we talk about the technology is, is really interesting. And Rogue One's the first time I think there's been a real freak out

about this kind of technology and what it means because they've also, cuz ethically, what's worth thinking about is that these two actors that are being replicating this movie have not given consent for their images to use because they're already. Yeah. Um, that, that is a really fascinating, like really fascinating, um, kind of ethical dilemma of like, who owns the rights to your image, like beyond all of the political and the, the nasties that you can do, there's that whole ethical concern of who owns your image rights.

M.K: Yeah. Cuz I remember watching that in the cinema and afterwards my mom was like, you know, that guy was dead. I was like, no. Cause I just didn't even notice. And it's the same thing with like, with the deep fake that I made. It's like I watch it and I think, yes, that's fake, but like when I just sort of preliminary showed it to people, they seem kind of like shocked by it cuz if you are not aware that deep fakes exist, you don't, your mind doesn't sort of see the possibility that it could be fakes. You just kind of assume it's real. Yeah, yeah. Yeah. So it was like the first one I did, it was like quite sort of glitchy because of like the way the mouth was moving. Exactly. But everyone's like, oh, I just thought it was a glitchy video rather than, yeah. It was something that like was completely made up,

P.T: Yeah, so what's worth having a look at, you know, is referencing popular culture. I mean, Rogue One was made in what I think 2017 or something. Yeah. Um, that's really the quality difference in four years. So when you talk about glitchy, the technology, some of that stuff, like even within three or four years mm-hmm. now, um, the difference between computer generated a deep fake generated algorithm generated like images, the, the quality difference is massive. So in 18 months time, that glitchy one that you've made would be almost indistinguishable. Yeah. And from your limited resource, not a Hollywood studio. Exactly.

M.K: Cuz I think the whole point that I'm making is that yes, you know, everyone knows CGI exists and everyone knows that if you see a dinosaur on tv, it's not gonna like be in real life.

P.T: Yeah, yeah, yeah. But if you see something believable, how do you know that? It's like not real. Now there's a, there's another couple of ones that I think are worth referencing as well, and these come back to some of these ideas. So there's, um, this whole idea of, you know, I've seen these putting people out of time. So, so that's a Tom Cruise deep fake. Yeah. Um, which is actually really, really good. Oh, I'll talk about this stuff in a minute. Yeah. Um, but the Tom Cruise one's a really interesting, because again, this is somebody, this is somebody kind of mucking about. Yeah. Producing content and spitting it out to demonstrate how easy it is to do this stuff. They're putting up on TikTok to make a point, but actually when you look at the Cruise video, um, it's, it's, it's really good. Like really, really good. Yeah. Um, and the reason it's really interesting, and sorry if I'm going too fast and need me to slow down,

M.K: No, no, no. It's fine, carry on.

P.T: Now the reason, it’s so good is because there’s such a large dataset of his face. So what you end up getting, so. Any kind of AI technology, deep or machine learning or otherwise, needs a substantial database of reference images or reference content in order to get really good. Um, There are, there are apps and things at the moment called like, I think they're called Vanna and the DALL-E where you can upload five versions of your face and it generates all these AI images from you. You're uploading five images and it's getting pretty good. Someone like Tom Cruise is easily replicable cuz there are literally hundreds of thousands of pictures of online. And what the algorithms are trained to do is not really generate the imagery because the, you know, the Hollywood and computer games industry have been doing that for a long

time. What the algorithms are really trained to do is, is fish and interpret the source material. Yeah. And when you have a very large source material, then you've got an algorithm that can really start to do some really like powerful stuff. And so there's, there are going to be people, uh, famous people who are probably more worried about this technology than. Other famous actors or not so famous actors, largely cuz there's way more source material for them to, to look from. So the, the Tom Cruise stuff is like really scary. Good, but largely cuz of the data source. Because like you've, like there are photos of concrete making like so many different expressions and so doing so many different things.

M.K: Yeah. Because one of the things I'm arguing as well is that because everyone's then posting photos of themselves online, far more now than previously, the technology becomes far more powerful.

P.T: Yeah, exactly. Now the other thing, the other thing, so, and I can talk to you about the tech in a minute. The other people that are most at risk are politicians. Yes. Um, you know, and politicians are most at risk. for two reasons. Um, one, there's a large amount of source material available cuz they're, they're public figures. Um, and there's video image and sound data. So you need all three of big reference databases of all three of those things to make a defect really good. So there's a lot of images, there's a lot of video, there's a lot of soundbites of, you know, prime ministers and presidents and things like that. Because they are who they are and they set policy, they set rules, they set the tone of, of media. They're probably, in all likelihood, the most, um, dangerous people to replicate online because they yield the most power. So the, the at the moment, and a good example is that at the moment, um, Zelensky is the most deep faked face on the planet.

M.K: Oh wow I didn’t know.

P.T: Um, he's one of the most notable public figures at the moment. Um, it was Barack Obama and I think it's now been taken over, um, by, by Zelensky. But yeah, he, he's got, he, there's a lot of propaganda that's been generated by, um, kind of non Ukrainian state. Um, I'm not saying who, but some of it's a bit wonky, but there's a lot of it going out at the moment. Yeah. Cause there's a lot of propaganda that's been coming out of, you know, all the various factions involved in what's going on. Um, so that, that's where, you know, but there is a lot of image material. He was an actor before he was a president, so again, he was quite famous actor. Yeah. On tv. So there's a lot of sourced material. So his, his deep fake stuff's gonna be like, pretty decent. Really, really good. Um, but that, that's where I think some of the, the big harms, you know, are gonna come from is this, you know, it's, it's not just the technology to replicate it, it's the training data available. Yeah. Um, and that, that I think is the really, that's the critical difference. The technology. You can do it, I can do it. You know, it's having enough source material to make it good. Um, the other thing that's changed a lot in the last 18, , um, which again is a big leap forward for deep fake technology, um, is, is real time synching. So what we've also seen, and this is coming out of Hollywood, so this is coming out of cuz of films like Avatar, um, which is, you know, James Cameron's big pioneer. Some of this technology in the original avatar was, um, real time lip syncing with, you know, a human and an avatar. very good. Deep fake avatar is another big leap forward. And so a camera being able to interpret your facial expressions and then mimic that in almost like that computer game spot I showed you at the beginning. Yeah. That's like another huge leap forward. And that's because camera technology's got much, much better. Yeah. Um, and so, so that's the other kind of facet to think about when you are writing is that, you know, we've all. Um, uh, a smartphone and the

cameras and those smartphones are almost as good picks on HD depth as, you know, professional camera equipment was five or six years ago. Um, so that, so being able to do things with cameras has now kind of massively amplified the opportunity for deep fake technology. Yeah, so really like, really kind of, it's the leaps in the bits of tech and the amount of data that's like the big shift change. It's like the really big shift change. Um, now, the, the other one. The other one. So I turn my cell is the other example I've got, and I'll show you how he did it, is the, when I, we were doing this talk, right. So yeah, I watched it. Yeah. So he's awesome. So my mate, um, we had a lot of footage. Like we had a lot of footage. Um, and, you know, he, we'd filmed, we'd started filming kind of a whole bunch of interviews for a documentary and things. So there was a lot of footage already available, and then sadly, he passed away. Now what we were able to do was, and this is how we made it work, so we were able to take the footage we already had and what the tech, what the tech does. is you are able, you literally just pin, pin that face like that guy did with Tom Cruise to the key parts of the face. Now this was stuff that was largely only available to Hollywood 10 years ago. We are doing it on a, on a modest computer setup in central London somewhere. We've got a, an algorithm that we've basically taught where the key pieces of the face are and the eyes and the mouth. So the most expressive pieces. Um, and then. The, the algorithm is then able from the photo or the machine is then able from the photographs and those key positions and his, his movements to basically replicate that. Now what we then did, so once we'd, once we taught the machine the key areas, so we'd effectively computerized this part of his face. Yeah. Um, the machine is able to kind of fairly seamlessly lay that on top of other video. . Um, so we reused, uh, yeah, we, yeah, we, we reused the same footage from earlier in the talk, if you like, of him speaking to the camera. Yeah. But what the computer's done is it's basically overlaid a computer simulated version of his face into the gap, into the space. Yeah. And then we are able to put his words through. through that graphic that then replicates his mouth movements and his head movements and everything. Yeah. That's so cool. So, so he'd written all of the copy that he was gonna say, but he just wasn't around unfortunately, to tell us what he was all, to read out what he was gonna say. Yeah. So that's how we were able to get the, the monologue at the end of the talk that I did or the, the stuff that he was going. So that's, . Yeah. So in the, in the, in the early parts, that's all the stuff that he had done with us. Yeah. Um, I'd been capturing quite a lot of what he was, he was thinking and saying, anyway, that was something else. Um, but yeah, this video at the end, any of the, and actually this is where I think a bit of trickery and magic comes in. You know, you were saying your video was a bit weird, like it wasn't quite perfect. Yeah, it was a bit glitchy. Um, he, he, and this is slight. Is where the video that we had at the end of the talk cuts to different shots, like there real footage. That's cuz the, the actual video that we'd made was a bit glitchy. Oh yeah. So we only used the voiceover goes all the way through, so we used the replicated voice. So the voice audio was actually pretty decent. . Interestingly, when you talk about how you spot deep fake, his mom, Leslie knew about this, right? Um, so she knew we were gonna do this for the talks, and she was totally on board and so was his dad. Um, they were in the room. Les was like blown away by the, the face. Um, but she was like, the voice was wrong. Oh, no one else knew, but mother would know that. So she could tell there was like something not quite seeking up with it, but. Yeah, we, we still couldn't get it perfect. There was still moments where it was a bit wonky and a bit glitchy, so we just cut it with other bits of footage.

M.K: That's quite clever.

P.T: Yeah. But yeah, all, all of this stuff is, this is all computer generated. That's so crazy. But again, the ethics of this is, you know, the ethics are, um, we have permission from his family. We did not have permission from James. Although knowing him, I could infer that perhaps he would've given us. But the ethics are, you know, we have a room, we have a thousand people watching some footage of somebody that they're deeply moved by. Yeah. Who's not actually said it. Like, yeah. What, what does it matter? I mean, it kind of does philosophically. Yeah. But I feel like the same message as being like sent across it from the people's point of view, it's the same. Yeah. That's the thing.

But it is, it is a, it is to a certain, from a certain perspective, a lie. Um, So, but yeah, really interesting. I'd actually put, no one spotted it, but I'd actually put a, a PS on the screen to say that some of the footage in the talk had been generated by the AI work that we'd done.

M.K: Wow.

P.T: Um, so there was like a caveat in there. No one seemed to have ever picked up on it, or, or questioned which bits it was, but that's how we did it. Um, but again, back to that original point that I was making about source material, um, you know, we had a lot of source material to. Yeah. Um, and that, that's the big difference. You've gotta have a lot to create something very realistic. Um, now, now there's something else going on at the moment, which again, in terms of deep fake and what I think philosophically and ethically is some stuff that's gonna start, I think making a big difference. Is, a couple of weeks ago I used a system called VA to play around with some of this generative ai. So, so all the deep fake stuff fits under the umbrella of, um, what we would refer to as generative ai. So that is a compute system, a tool, uh, set of algorithms that can create brand new material from all sources. Yeah, brand new. So it's not, it's not like, um, process. old stuff and reusing. It's generating stuff that's completely new. Um, and, uh, DeepMind, Google DeepMind were kind of really one of the pioneers of generative technology. Um, yeah. Amongst other people, but. Um, one of the things that happened a couple last year was that lots of these, um, startups started releasing their tools for people to play with, and it's one called VA Now VA's really fascinating. So it generates all of these weird portraits. It makes you upload five pictures and it kind of yeah, creates fi you know, hundred pictures of you doing really weird stuff or whatever. I submitted my pictures of me without. Yeah, so my photos go up five photos without my glasses. I'm completely by 'em without them, but I just thought, I, I thought that might cause a problem. So it goes up, um, 30 minutes later I get an email, I click in, there's my gallery. They're all kind of hilarious. Me and my wife are laughing at them and I'm like, oh, they've put, picked, they've put glasses on some of these. Yeah. And not just any pair of glasses, like the glasses that I wear on some of these pictures.

M.K: Um, that's really weird. How have they done that?

P.T: Now what they've done, and this is where I think there's also some ethical concerns that we all need to be careful of, and we can talk about the politics and stuff as well. But, um, what they seem to have done is fill in the blanks in their training data by having their algorithm go off to like Google images and have the algorithm go, well, that's Pete's chin, nose, eye. So the stuff that we do manually to James. Yeah. Um, and then use that kind of, um, almost surveillance type technology to find other pictures that are 95% plus, um, certain to be me. And then filling in the gaps in their training data with stuff that they find online. And, um,

most of the images on Google, images of me with my glasses, um, because I wear them all the time. So it's, I've not given it permission to go off and look for other pictures of me, but it has gone off across the internet and filled in its blank. Yeah.

M.K: So if I did that without my glasses, I assume it wouldn't find photos me with the glasses, so that would be interesting to see if it did. Yeah, because there are some photos of me, like on my Instagram, whatever Yeah. Things that they shouldn't have access to.

P.T: Yeah cuz there's, there's a lot of publicly available pictures of me doing talks and doing professional things. So that's why it's got me worn my glasses. Because the other interesting thing is, I figured how, how it was doing it, because I wear two types of glasses. Like home glasses and then professional ones. So it helps me kind of swap modes a little bit. So I only ever wear these kind of when I'm in work mode. Yeah. And that's what all the pictures online are with me, with these in work mode. So, so that like, so the re the other, so one of the questions or one of the things I pointed out at the beginning was that, you know, AI needs, um, algorithms, need neural networks, need a lot of training data to make stuff. It, the big advancement has been in almost that surveillance approach to training data, which is, if I have a small amount of data, how do I, how do I create the data that I need to fill in those blanks? And that's why going off and searching the internet or going off and, um, you know, creating references from other places and then pulling that back in. That's why Tom Cruise can be so easily replicable. It's why Eric Baer and stuff like that. But they're now, they've worked out how to do it to. So it starts to become really kind of muddy now. Starts becoming more complicated.

M.K: Yeah. Like with my head mistress, I had to, I, I went to her like set up a meeting. I was like, Hey, this is what deep fake are, this is what I wanna do. Do you allow me? She was like, yes. And then like I went again to get with the permission separately for like the audio and the voice. Um, it's the audio and the video. And um, and then when I showed her what I had at what. Um, in another meeting, she, she was quite sort of shocked by it. And even if that was like the glitchy version. Yeah. And I think cuz when you see it of yourself, that's when it gets like more scary. Yeah. And I think she then, yeah, she then said to me, she was like, you are only using this for this part first. And I went, yep, of course. Yeah. And she said, you will delete it all in front of me afterwards. Cuz that's what, just evidently like how like scared she is, cause I said, oh, I could do an assembly on this if you wanted. She said, I'm not sure if you do that, then even though I know you won't use it for malicious purposes, someone else in the school, probably might. Exactly.

P.T: And that's the thing Now, that's, that's where you get to the stuff right now that I think is gonna be really ethically quite complicated. So, you know, um, because you know, not only can you start to, um, replicate people to do some real fun stuff, you know, jokes and memes and or, you know, dead people. Um, be it actors or, or people that aren't with us or whatever, like James. But then you start to get into kind of, you know, um, you know, deep fake porn is huge for celebrities at the moment. Um, like really horrible. Um, there's kind of, you know, deep fake political speeches and statements, which you've already mentioned that's really horrible. Then there's the bullying and the trolling and, and all that kind of stuff. Um, some of the, some of the most dangerous deep fake stuff, um, which I found online years ago, a few years. Uh, when it's, it's shot. Um, uh, so I mean, we've looked at a few examples that are sort of crystal clear and find a good example. We've looked at examples that are like, you know, crystal clear and, um, like really good quality. Yeah. Um, some of the, some of the stuff that's

online, like deep fake wise are when, I mean, this is one that we. it was, we were looking at and it was, you know, um, I can't, I can't remember exactly what it was. I think somebody, one of these cheerleaders Yeah. Is, is like, not a cheerleader, it's somebody else or something. Yeah. And then somebody's used it to basically spoof an exam or, or, you know, X. It's like to, yeah, like put it on your cv. You're like, oh look, here's a video of me doing X. Yeah, this is what I did now. But the point was this is really easy to do. Cause the quality's bad. Yeah. So there's this whole like really interesting area of like, okay, you can have stuff that's really high quality that looks indistinguishable, but then actually what we're probably gonna see online, or a lot more really bad quality videos that could be interpreted as being somebody. Um, be that, you know, fake CCTV footage or, you know, things like this video or whatever that are actually such bad quality that the fact that the technology's not quite, um, you know, up to speed from people's phones and bedrooms, whatever, is actually playing to the advantage. So the, the really kind of weird, sort of deep fake, like the fringe deep fake stuff, the really damaging stuff is the, is probably gonna be the poor quality.

M.K: Yeah, I haven't thought about that.

P.T: Yeah. Not the, not the high quality. Yeah. Because you'd think it has to be the high quality stuff to make it believable, but equally so, like when I did my lurk and video and people just assumed it was just a bad quality video. Just believed it anyway. Yeah, exactly. So the, so the like, and that's, that's where I think, um, yeah, that's where I think there's gonna be, there's this kind of very muddy divide. Um, now the. The other bit that I, if I was you. So there's faces and the implications of, you know, who owns the rights and permission and the damage you can do with good and bad quality. Um, voice, I think voice biometric is, is a really dangerous space. So, um, voice biometric is used in an awful lot of security settings. So voice biometric is used for kind of, you know, um, Uh, on banking systems, they quite often have a recording. You know, when you go onto a bank, you probably don't do it, but I'm old, so I'm used to going onto the telephone and them saying, can we record your, can we record the conversation for training and monitoring purposes? And you go, yeah, that's fine. What they're also doing is having your audio on file so that they can then start to use systems like Darktrace, I think as one to really issue, like really be assured that when you ring up, you are who you. . Now the problem with deep fake technologies, it's getting really, really good at replicating people's voice. Biometric. Yeah. And therefore, being able to replicate people on the telephone to very convincingly hack through security systems. Or have, you know, your boss agree to something that they wouldn't ordinarily agree to or say something. Or, you know, um, old people may have their, you know, quote unquote children ringing up. Can you just transfer me 200 pounds please, mum, cuz you know, I've gotta pay the gas bill, whatever. But it's like, just really good deep fake technology for scammers voice, I think is more dangerous because it doesn't have the problem of visual. Yeah. Um, it has the, again, it's like a posit, it's like the low quality video example I gave you. Um, voice is great because you are only assessing it with your ears, not with your eyes. Um, and I think, so I think you're talking about like dangers and harms. I think spoofing your head mistress and having her kind of freak out a little bit or your teachers freak out a little bit. It's kind of funny. Um, her voice can be 10 times more harmful if that's replicated in a really kind of quality way.

M.K: Yeah. Like the voice is not very quality. It's a bit robotic. Yeah, but I think, I think it always works. It, it works that it's not very good. I mean, I can show it to you if you like actually.

P.T: Yeah, please. What did you use?

M.K: So I used, um, there was this Wav2Lip tool for like the lip syncing. And then I used Descript for the voice.

P.T: Amazing.

Video: “Um, I just would like to make a quick announcement, which will go up to your parents later tonight…”

P.T: That’s really good!

Video: Which is that school next year will begin at 10 in the morning. We have come to this decision following scientific research, which has shown that students benefit from a later start of the day. However, this does not mean that we will lose school time as we will also be going to school every Saturday.

P.T: Yeah. See, I got it. Cause I know what I'm looking for. I got like, the tongue is too red and the lips don't quite sync up, but that's really good. Yeah. Um, considering you’re, no disrespect, a student and you're playing with this stuff.

M.K: Yeah. And it was from one video that they got this as well.

P.T: Yeah. The fact that you are able to do that already, like I wasn't even expecting to see you being able to do something that good with so little source material. So even I'm quite shocked by the source material.

M.K: Yeah it was just this one minute long video that I fed it.

P.T: That's, that's really incredible. Like, that's blown me away because, you know, I think I, I gravitate, I'm in a world of, you know, I've got a 13 year old son who's mucking around with stuff and I've got a professional career where we're trying to generate sometimes good stuff in the ethics, but, A student mucking around with it for a project and you are able to do that. Like, that's a leap.

M.K: I mean, so people at our school can probably tell the voice is a bit weird because it sounds a bit too robotic. And she's a very sort of emotive person when she talks. Um, but even I found it goes through, I just did this first try and, um, I found that the, her movement sort of matched the kind of what she was saying. And that was, and I just saw that and then I was like, okay, I'm not changing the words or anything cause I'm gonna leave it like this.

P.T: It's really good. And that's, and that comes right back to that thing. So I showed you the avatar at the start and we put the text through it and it kind of, you know, says whatever you want. Like you're doing that with video, um, crunch that down to a lower quality like mobile phone quality that goes viral. Yeah. And the, the fact that it would be low quality, you'd get away with it. Yeah.

M.K: Cause initially I think of changing it because it looks like she's being interviewed. So I could pretend it was like a recent interview. But if anything, so like, if you really think about it and what I'm gonna explain in my talk. You see this? This isn't our school. She moved to be head of our school a couple years ago. This is from previously. Oh, this is a couple years old. So I, I can easily say look like our school, our science lab don't look like this. This is from Wimbledon High. Yeah. Um. . So anyone who's like really paying attention Yeah. Can easily

tell. But it's the sort of thing that you don't have that voice in your back of your mind saying, this might be fake. You just assume it's real.

P.T: So the, so one of the big areas, and again, I think it'd be worth sort of googling around and researching this a little bit. So I know a lady really, really well, learned a huge amount from a lady called Jenny Radcliffe, and she's a, uh, social engineer. So, um, they're incredible, like, um, professional con artists. Yeah. So their job is to, to. , um, businesses to show where their security weaknesses and things are. Um, she, she was telling me about, you know, the, the security implications of, because of stuff like this, but also for banks and, and things like that. Like the social engineering will be using deep defect technology quite liberally.

M.K: Yeah I was gonna write about that.

P.T: Yeah. Yeah, look for Jenny Radcliffe. She love loads of stuff online, um, interviews and things like that, but yeah, so the social engineering and deep fakes can be really bad.

M.K: Yeah, cuz I thought as well, cause even now we have like phishing emails and things like that, but imagine if instead of a phishing email, it was a phone call from a friend.

P.T: Yeah, that's it. Or Snapchat or whatever. It's just like once you start getting these kind of like, fast moving videos that go up and down. TikTok is another one that I think deep fake is just like ripe for, cuz. Short burst, low-ish, res, you know, phone size videos, um, cut as memes, you know, overlaid of copy. So people are never quite a hundred percent paying attention to the video. They're sometimes looking at the caption or they're sometimes looking at, you know, whatever else is going in the music. And so the, the platforms that are coming out for youth, um, lend themselves really well to deep fake. You know, banks and security systems are absolutely ripe for deep fake because they're, you know, flawed and they use voice biome, things like that. So, um, I don't, I think the next couple years we're need to see like a per a proliferation of this kind of technology going all the way through. Yeah. Um, but, and also don't, I mean, don't discount, so that's video. I'm talking about voice. I think the other one that is just gonna be in separate, it's just gonna go crazy, is just images. Yeah. Images of people. Generated in precarious situations or, um, you know, in moments that can, uh, damage them or whatever. Um, I think it's gonna be really kind of big. Like that's gonna be a big, big one. Yeah. For trolling and for things like that. Yeah, but I think what I've, one of the problems with images as well is that people are familiar with Photoshop and that's why I think that sometimes they might have less sort of like power.

M.K: Cuz if I'm like scrolling on Twitter or something, I know that a lot of the images I see might not actually be real, but people don't really think that about videos because you think, oh, you can't Photoshop like a thousand different images and put them into a video.

P.T: Yeah, true. Yeah. True. There's, um, I mean the, the, I guess the, I guess the other thing that's happened over the last five or six years, And it comes back to that point I'm made about training data is, you know, probably 10, 15 years now actually is as humanity, we are generating a huge amount of a certain type of data and that is imagery data. So, um, you know, you can track a lot of this back to, so Facebook in its 2005 or whenever it was, it came out, was collecting a vast amount of people's data when they were uploading. Mm-hmm. , quite cynically what they did, quite cleverly actually. It was cynical, but it was very clever. When Facebook first came out, what they wanted to do was kind of be able to tag you another photo. So they wanted to know, you know, they wanted to say to my mate, Neil, is that Pete

Trainor in your picture? So they could create the social graph. Yeah. What they, what they needed though were, was training day. So Facebook actually sponsored, um, or invested quite heavily in companies in China that we were creating selfie. Yes. So, because the point of a selfie stick is you get long range and there's normally five or six of you or two of you in a picture. Yeah. So if you encourage people to upload photos, well there's more faces than just one. You're doubling the amount of training data and within about six or seven months of investing in companies that, that are creating selfie sticks, Facebook, loved it. The amount of volume of biometric face data that was being uploaded to the platform. Um, and therefore, if you like tracking those people, tracking people becomes infinitely easier cuz you've got a much bigger data set and that, so that also comes back to my point about images. Um, images were easy to replicate a long time ago because there's a lot of facial data of being uploaded cuz of things like Facebook. Um, I mean, it's literally called Facebook now. We've got TikTok and Snapchat and stuff, and people are pushing videos up there. We're generating a lot more video based content and therefore video based, um, you know, deep fakes are becoming easier to, to replicate cause there's more video source data for it to reference. So it all comes back to the sources of data day. Yeah, yeah. The tech has been there for a while. Uh, the quality of the output is, is all entirely driven by the data.

M.K: Do you have anything to comment on, how we can sort of protect ourselves against the technology?

P.T: Yeah, so, um, so one thing that is really interesting, and again, it's a parallel example. Um, so chat G P T comes out at the end of last year. Yeah. It's really, really good. Even I'm impressed with it. I'm as impressed with it as I was with that video you just showed me. And I'm gonna, an area domain expert, if you like. Um, chaps, GPT comes out really. Really, really good at generating and creating content. So another generative example, um, but it's had a huge amount of source data. Other people are starting to create algorithms and tools to spot G P t generated content for schools so that they can monitor essays that are submitted, um, or whatever. There are tools emerging using. Deep fake technology to spot deep fakes. So it's turning into like an arms race. So they're creating software to spot software. Yeah. So one of the things that's gonna protect us is government and the security services will be investing really heavily in like darktrace type technology to spot deep fakes. And it'll be things like, um, there'll be, at the moment there'll be like little cues or little things that an. Would pick up on or voice biometric or whatever. Yeah. Tiny little things that, that only experts can spot, but software will be able to do it really, really quickly. So that's the big thing that's gonna protect us from this. Mm-hmm. . Um, I guess the other thing that's gonna protect us from this is having everybody educated enough, um, to ask and question, you know, to question things that they see. If it seems too good to be true, it probably. It's, um, but again, that's a social problem and it's not something we're gonna solve really, really quickly. Um, uh, it would be having people not uploading their images into these machines to like generate stuff is like another really interesting one. Um, but, you know, again, how do we stop that kind of stuff happening really difficult. Um, incidentally, Facebook again, semi responsible for a large uptake in video content, uh, because they instigated the Ice Bucket Challenge. Do you remember that one? Yeah. So the Ice Bucket Challenge was years ago, whatever. And what they did was they encouraged, they encouraged, they wanted to get an uptick in video data that was uploaded, so they created a campaign, the ice bucket Challenge. Um, so lots and lots and lots of people were uploading their video and then using somebody else's name. So they were able to string together. We were all encouraged to upload video content, and we just did

it like sheep Following the Sheep. Yeah. I'd never realized that before. Yeah. Boom. So they, they kind of led us to the, to the data, the slew of data that was created and we all did it. What we have to do, start educating people from a school level that their data is being used and manipulated for these things. So that they are, they are uploading it or they're giving it away with some at least knowledge that it can be abused. And I think, I think that's the only other thing we can really do to start to protect people is make them aware of the fact that, you know, content is replicable if you give it enough source material. Um, so that's the other one. Another thing that's worth mentioning by the way, is just backtracking a. , apparently there was, I have not seen any research practice up, but I'm hearing it. Apparently there's significantly more, uh, male deep fakes than female deep fakes. Cause men tend to have shorter hair. Because the other thing that's really difficult to replicate on video is hair. Yeah. So your hair goes over your eyes. Yeah. You know, doing that, you'll be able to get your mouth. But kind of all this around there is kind of covered.

M.K: And even if you then move your head, how's the hair like?

P.T: Yeah. Yeah. So hair has always been really, really difficult. Even in, even in animation, even in um, top quality, CGI hair has always been one of the things that kind of creates that uncanny valley effect. Yeah. Um, so that's why there's a lot more, apparently there's a lot more deep fake of, of short head men and short head women, but maybe short men than a, there are other areas cuz they're, you are literally cutting out and so, um, you know, again, You can, you can start to spot some deep fakes because there are areas of the face that are just really, really difficult to replicate, like hair.

M.K: I've, no, I'd never thought about that, but just thinking about how I, sort of like making people aware that their, like sort of their data can then be used to replicate their content. What was interesting is when I did the Descript and I used like the audio files of my Head, I went to her and before you submit all the training data, you have to get an extra thing of that speaking, hi, this is Mrs Lunnon. I permit my voice to be used. I understand that this is happening, da da da. And so you have to have that person speaking give their permission. Yeah, in order for you to be able to use the software, and I thought that was really good as well.

P.T: Yeah, that's really good. Um, the best voice, there's something called poly.ai. Poly poly ai, um, was getting really good at using what you just asked your head teacher to do to replicate the voice. So again, that's another one. So when you get people to, so poly is a system whereby, you know, you would talk four or five statements. Follow it on screen or whatever, or agreed something and it was using, it would use that voice biometric to replicate your voice. So, you know, having her agree to that stuff ethically is great. But you've also just created a really interesting part of source material cuz people tend to speak clearly when they're asked. And they're prompted to answer a question. Yeah.

M.K: And it was good cause all the source material that I had of her for that voice was her from doing things like interviews, like what I wanted her voice to be because I tried it with my voice. But I was, uh, reading this random document at like seven in the morning, and then when I, when the voice was like, properly cloned, I could tell that it was like a really tired version of my voice.

P.T: Yeah, yeah, yeah. Um, the, the other one, just to bear in mind on the voice thing, um, is accents are currently really difficult.

M.K: Oh, I heard, yeah.

P.T: I heard in a case study from a voice activated lift in Scotland that just didn't work or it had like a 70% fail rate cuz it just didn't understand people's accents. Um, and Alexa is really bad at understanding people's accents and stuff, so, oh, I've seen things wrong, right? Machine interpretation of accents is like still really, really bad. So, um, like posh white people, posh white men are basically, we're the, we're the most replicable, deep fakes on the planet at the moment, I imagine. Yeah. Cause we don't have long hair and we have relatively monotone accents, so it's like we're in trouble. Um, but yeah, there, there's, there are things that I think keep certain groups of people relatively safe, like those, like difficult to replicate things, but you know, actually I think we just have to educate people. The fact that this stuff exists and it's app based and it's mainstream and it's easy to do, easy-ish to do, um, yeah, it's not going away. That's gonna get better as well. Yeah, much, much better.

M.K: Thank you so much!

P.T: You're welcome. What I'll be able to do, um, is I can send you those videos as well.

M.K: Thank you so much.

P.T: I'll upload them Dropbox and send you that link on your, that link so you can kind of use whatever ones you want to kind of have a look at, um, or use. Um, and if you have any other questions, you can just email me. Alright. Cause I'm quite happy to answer them. But I hope that was, I hope that was semi useful.

M.K: Yes that was so useful. Thank you so much.