

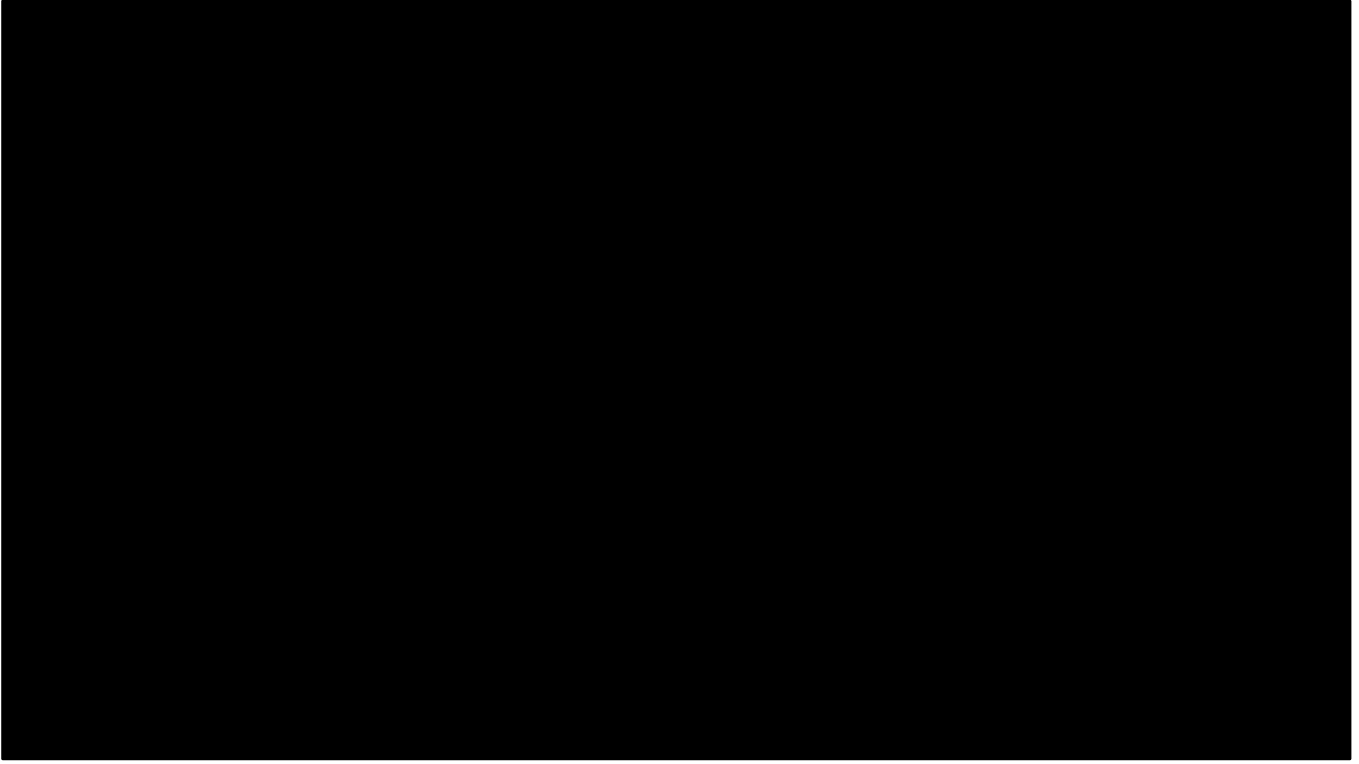


Unmasking Deception

Navigating the Era of Deepfakes and Misleading Content

Hello everyone. I'm Phil Hood, learning technologist at East Riding College. I was asked to do a presentation on pretty much whatever I liked but rather than discuss something that's strictly based in academia, I thought it might be more beneficial to look at something that has repercussions within education but also in our everyday lives, so I decided it'd be worth looking at the kinds of ways misinformation can be spread online, and the technology behind it.

Now before I get into the presentation, I've been sent a recording over from America that I'm going to play for you now. I've not had chance to watch it myself, so I hope it's appropriate...



Forms of Misleading Digital Content

Okay... Thank you current and former Presidents of the United States for your very kind words.

So to start with, I'll just run through some examples of the different forms of misleading digital content

Photo Manipulation



First up, is photo manipulation. Altering images to misrepresent reality. It can include techniques like...



Airbrushing

Airbrushing, to remove wrinkles and blemishes in order to create a more perfect image. It's kind of sad that in looking for a picture to demonstrate this, there were tonnes of examples of women, but barely any decent quality examples of a man.

Adding/Removing Elements



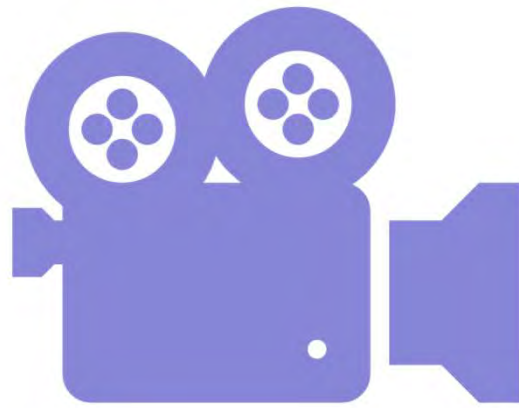
As basic as it sounds. Some historical examples, the one on the left Nikolai Yezhov was removed from this photo taken at Moscow Canal. And the image on the right purports to show the incoming disaster that was the 2004 Tsunami in Asia but was revealed to have been taken on the coast of Chile with the enormous waves simply added in via photoshop.

Combining Images



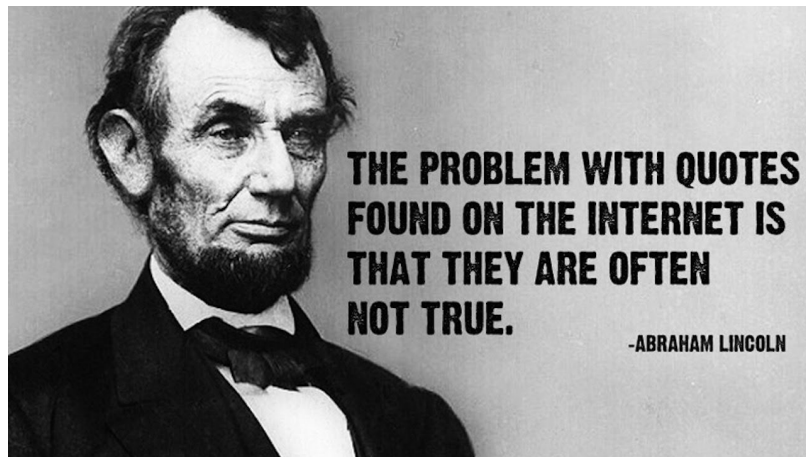
Combining images to create a false narrative, so here this impressive rice wave on the left image was actually the sculpture seen on the right that was edited into the wok.

Video Manipulation



Video manipulation involves altering or creating videos to deceive viewers and uses a lot of the same techniques covered in photo manipulation, only as a video.

Misleading Quotes



Sharing quotes out of context or attributing false statements to individuals can easily manipulate public perception. Quotes can be edited, misattributed, or taken from different contexts to distort their original meaning.

Satire and Parody

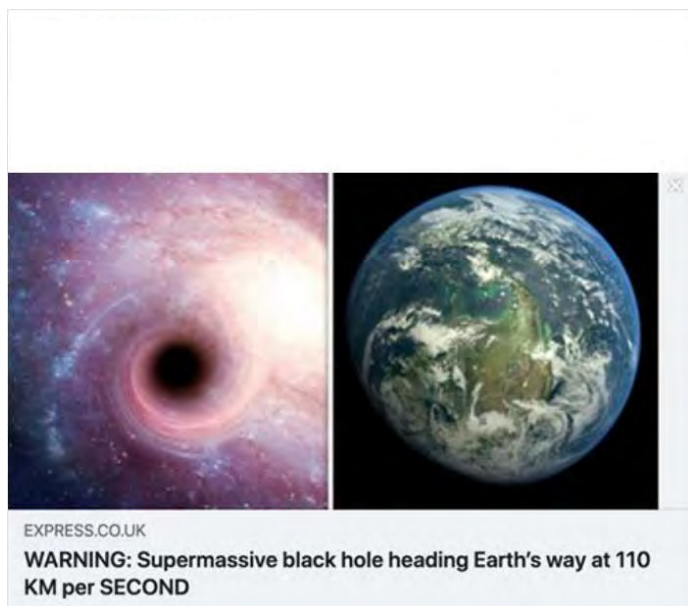
- The Onion
- Clickhole
- News Thump
- Private Eye
- The Daily Mash
- Babylon Bee



While satire and parody serve as forms of creative expression, they can sometimes be mistaken as factual information. Misunderstanding satire or sharing it out of context can contribute to the spread of misinformation, and it can make you look a bit stupid like this person on twitter.

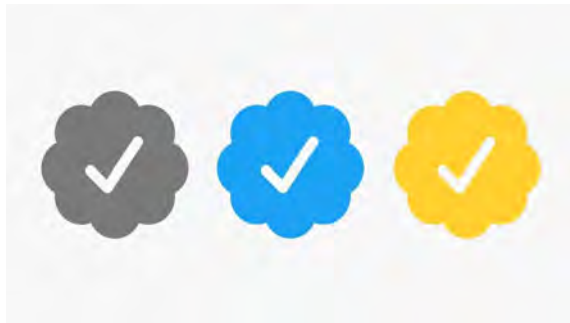
Some of the well known satire and parody publications would be the onion, clickhole, news thump, private eye, the daily mash, Babylon bee but there're loads out there.

Misleading Headlines and Clickbait



Sensationalized headlines or clickbait titles can draw attention and entice individuals to click on or share content without fully understanding its accuracy or credibility. These tactics often prioritize engagement over providing accurate information. So in this example, the headline is very alarmist but within the article however it mentions it will actually take 4 billion years to reach us.

Fake Social Media Accounts



Impersonating individuals or creating fake accounts on social media platforms is another form of misleading digital content. These accounts can be used to spread false information, propaganda, or engage in deceptive practices.

Legitimate accounts for public figures or organisations are usually identified via some kind of check mark next to their name that confirms it's official. However, since the incredibly insecure Elon

Musk took over twitter, it's harder as people are able to pay to have the blue checkmark on their account and others are seemingly given out without proper verification.

For example, the gold one, which is supposed to be given to legitimate businesses, was awarded to this fake crypto currency account on the right. I mean crypto seems like a Ponzi scheme to me anyway. But this is literally a fraudulent scam given legitimacy by multibillion dollar company.



Era of Deepfakes

Deep learning + **fake** events = Deepfake

So lets have a look at deepfakes, the 21st century's answer to Photoshopping. deepfakes use a form of artificial intelligence called deep learning to make images of fake events, hence the name deepfake.

What are they mainly for?

15000 deepfake videos Sept 2019

96% Pornographic

99% mapped *female* celebrities onto porn actresses



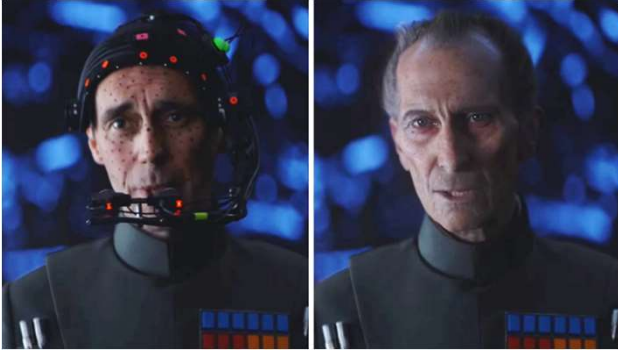
As with a lot of advancements in technology, especially where the internet is concerned, deep fakes have been used mainly for pornography.

The AI firm Deeptech found 15,000 deepfake videos online in September 2019.

96% of those were pornographic, and 99% of those mapped faces from FEMALE celebrities onto porn actresses.

In this more harmless example, Jack Nicholson in the Shining on the right has been replaced with Jim Carrey's face on the left.

What technology do you need?



Guy Henry becomes Peter Cushing in *Rogue One* (2016)



Robert De Niro de-aged in *The Irishman* (2019)

It wasn't long ago that to achieve this kind of effect, you would've needed a blockbuster sized budget to pull it off. For example, in the Star Wars film *Rogue One*, they used performance capture technology on actor Guy Henry to make him look like a 1970s Peter Cushing.

And in Martin Scorsese's *The Irishman* they developed what was essentially a camera with multiple and infrared lenses in order to be able to de-age the actors without having dots all over their face. All very expensive.

But the times they are a-changin'...

Originally required:

- Powerful computers
 - Reduces processing time to a matter of hours
- Expertise to reduce video defects

But now...

- Companies will do it for you
- Mobile Apps easily available

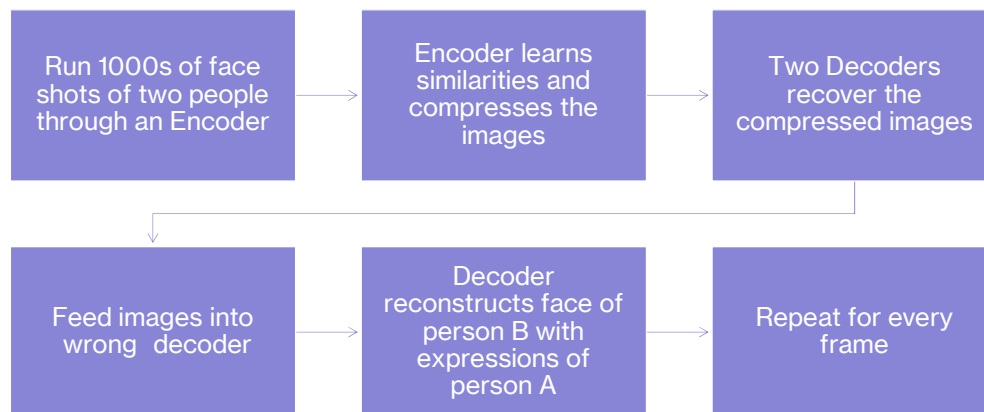


But technology is constantly advancing and becoming more accessible. So originally you would just need a powerful computer, which would reduce the time it takes to process deepfakes from days if not weeks, to hours.

And then it required some level of expertise too, to touch up completed videos and reduce any flickering and defects.

But now not only are the system requirements of a computer to be able to create a deepfake diminishing, and the technology becoming more accurate and believable, but you also have companies like *Deepfakes Web* which will make them for you.

Also there're mobile apps that will do it too, such as *Zao* which lets users add their faces to a list of TV and movie characters on which the system's trained.



So how exactly are they made? This hopefully isn't too complicated of an explanation, but it takes a few steps to make a face-swap video.

First, you run thousands of face shots of the two people through an AI algorithm called an encoder.

The encoder finds and learns similarities between the two faces, and reduces them to their shared common features, compressing the images in the process.

A second AI algorithm called a decoder is then taught to recover the faces from the compressed images. Because the faces are different, you train one decoder to recover the first person's face, and another decoder to recover the second person's face.

To perform the face swap, you simply feed encoded images into the "wrong" decoder. For example, a compressed image of person A's face is fed into the decoder trained on person B. The decoder then reconstructs the face of person B with the expressions and orientation of face A.

For a convincing video, this has to be done on every frame. Most films are have 24 frames a second, YouTube videos tend to have 30 or 60 frames per second.

or you can GAN it!

- **G**enerative **A**dversarial **N**etwork
- Generator turns random noise into synthetic image
- Synthetic image added to real ones, fed into Discriminator
- Repeat process until image becomes realistic



These people do not exist

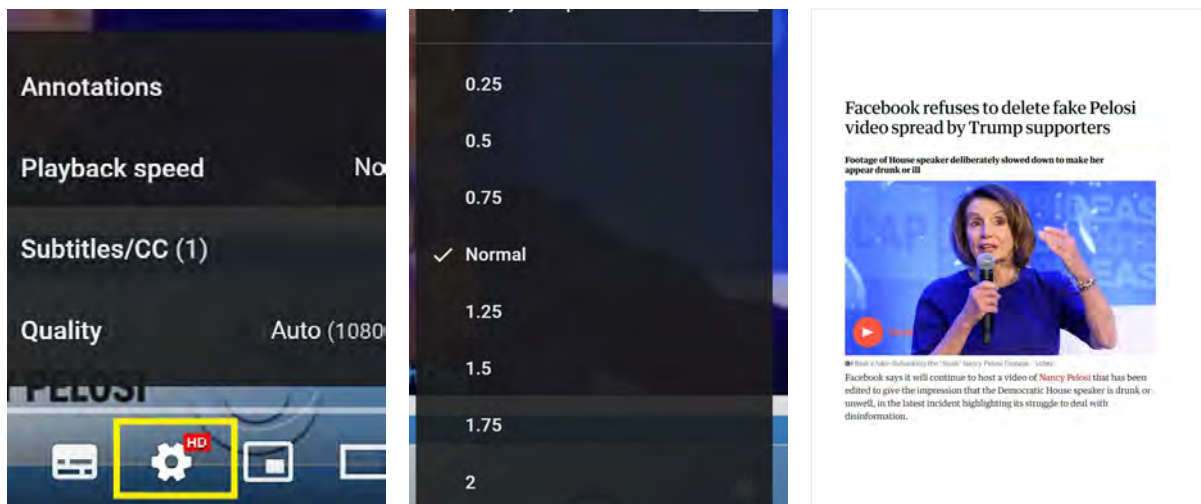
Another way to make deepfakes uses what's called a generative adversarial network, or GAN. I'll keep this brief but basically two AI algorithms are pitted against each other.

The first algorithm, known as the generator, is fed random noise and turns it into an image.

That synthetic image is then added to streams of real ones, for example of celebrities, that are fed into the second algorithm, called the discriminator. At first, the synthetic images will look nothing like faces.

But repeat the process countless times, with feedback on performance, and the discriminator and generator both improve. Given enough cycles and feedback, the generator will start producing **very realistic faces** of completely non-existent people, such as them on the right.

Shallowfakes



I'll very quickly touch on Shallowfakes which are videos that are presented out of context or are doctored with very simple editing tools.

You can easily try this yourself: go to any YouTube Video, click settings and then playback speed, and then slow it to three-quarter speed and anyone speaking in the video will suddenly sound like they're drunk. Which is what happened to US politician Nancy Pelosi when a slowed down video of her made the rounds on social media supposedly showing her slurring her words.

Not Just Videos



So this technology isn't just for videos. Deepfake technology can create convincing but entirely fictional photos from scratch. A non-existent journalist, "Maisy Kinsley", who had a profile on LinkedIn and Twitter, was a deepfake, likely created with a GAN.

Another LinkedIn fake, "Katie Jones", claimed to work at the Center for Strategic and International Studies, but is thought to be a deepfake created for a foreign spying operation.

SHARE

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies



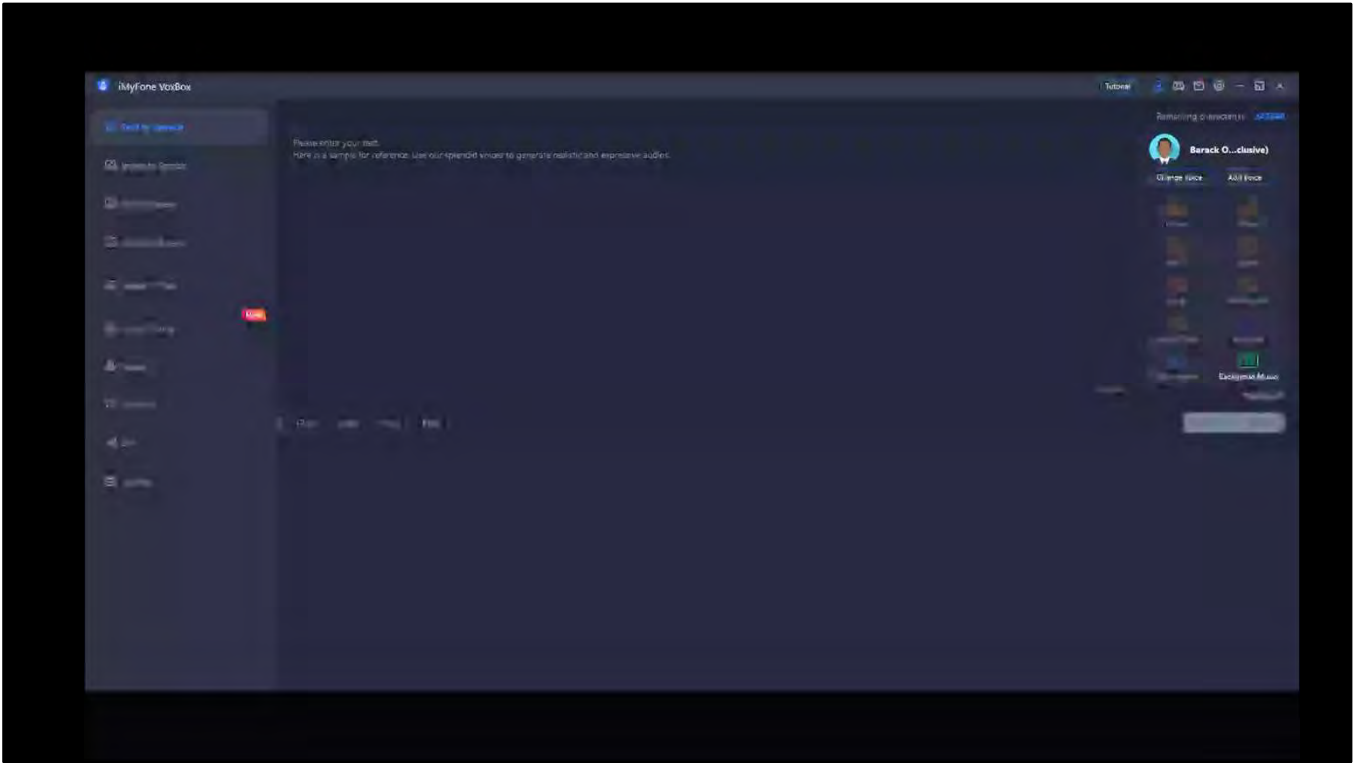
Sounds a bit fishy...

I have a confession...



I have a confession to make... That video I played at the beginning, which was sent to me by the united states government, was not entirely legitimate... I actually used a programme called VoxBox to insert Barack Obama's voice at the end of the video, it was not there in the original recording they sent me. Shocking I know.

I'll show you a clip of just how easily it can be done in VoxBox...



So, you put in your text, press convert, wait a couple of seconds and then there it is.

Implications

- Digital content depicting someone doing or saying something they did not
- Spreading of conspiracy theories
- Genuine content being denied as fake

“The problem may not be so much the faked reality as the fact that real reality becomes plausibly deniable” – Prof Lilian Edwards

- False evidence entered in court
- Mimic biometric data
- Potential for scams



Quick as that, though it did cost £12, and that was the cheapest option.

Anyway, let's have a look at some of the implications of this...

The most obvious one of course is videos, images or audio depicting someone doing or saying something they did not, and being able to easily spread conspiracy theories to people who are less technologically literate. But something else to consider is genuine things being denied as fake.

Lilian Edwards, an expert in internet law at Newcastle University said it best that the problem may not be so much the faked reality as the fact that real reality becomes plausibly deniable. For example, Prince Andrew denying the picture of him and an alleged trafficking victim was fake in his infamous interview with Emily Maitlis.

Realistic footage being entered as evidence in court cases, which is becoming an increasing concern in child custody battles and employment tribunals.

Mimicking biometric data. How many of you unlock your phone or authenticate for bank transactions using your face to confirm it's you?

Linked to that, a clear potential for scams. If someone telephones you out the blue and asks you to send them money, you probably wouldn't give them it. But what if you get a video call through WhatsApp and the person on the other end looks and sounds identical to one of your parents or kids?

What's being done about it?

- AI being created to detect deepfakes
- Digital ledger system holds records of origins of pictures, videos, etc
- Facebook ban misinformation produced using AI (but not Shallowfakes)
- The Law:

"Under a planned amendment to the Online Safety Bill, people who share so-called 'deepfakes' – explicit images or videos which have been manipulated to look like someone without their consent – will be among those to be specifically criminalised for the first time and face potential time behind bars." – 25/11/2022



So, what's being done about deepfakes and the harm they could cause?

Funnily AI already helps to spot fake videos, which is great but they work best for celebrities, because they can train on hours of freely available footage. But detection systems that aim to flag up fakes whenever they appear are being worked on.

Another strategy uses a digital ledger, called the Blockchain, to hold tamper-proof records of videos, pictures, audio and so on so their origins and any manipulations can be checked.

Facebook has a policy to ban any deepfake videos that are likely to mislead views, but they're not applying that to shallowfakes like that Nancy Pelosi video I mentioned earlier, which are still allowed on the platform.

In November last year the government announced new laws to better protect victims from abuse of intimate images. In an amendment to the Online Safety Bill, people who share deepfakes that contain explicit content will be among those to be specifically criminalised.

On the brighter side...

- Historical and cultural re-enactments
- Language learning
- Science simulations and experiments
- Accessibility and assistive technologies
- Film restoration and preservation
- Dubbing and Localisation



Now I'll be honest, I made this presentation and then realised I hadn't actually included any positive uses for deepfakes, and the attention its gained has been mostly negative, but there are ways it can be used for good, especially within education. Here're a few examples...

They can bring to life historical figures for students to interact with. Hopefully allowing them to better understand historical events, cultural contexts, and important figures in a more engaging and memorable manner. So here we have Harriet Tubman brought to life through the MyHeritage app.

Deepfakes can also help students practice language skills by generating realistic conversations with native speakers, allowing learners to improve pronunciation, fluency, and comprehension in a more authentic way.

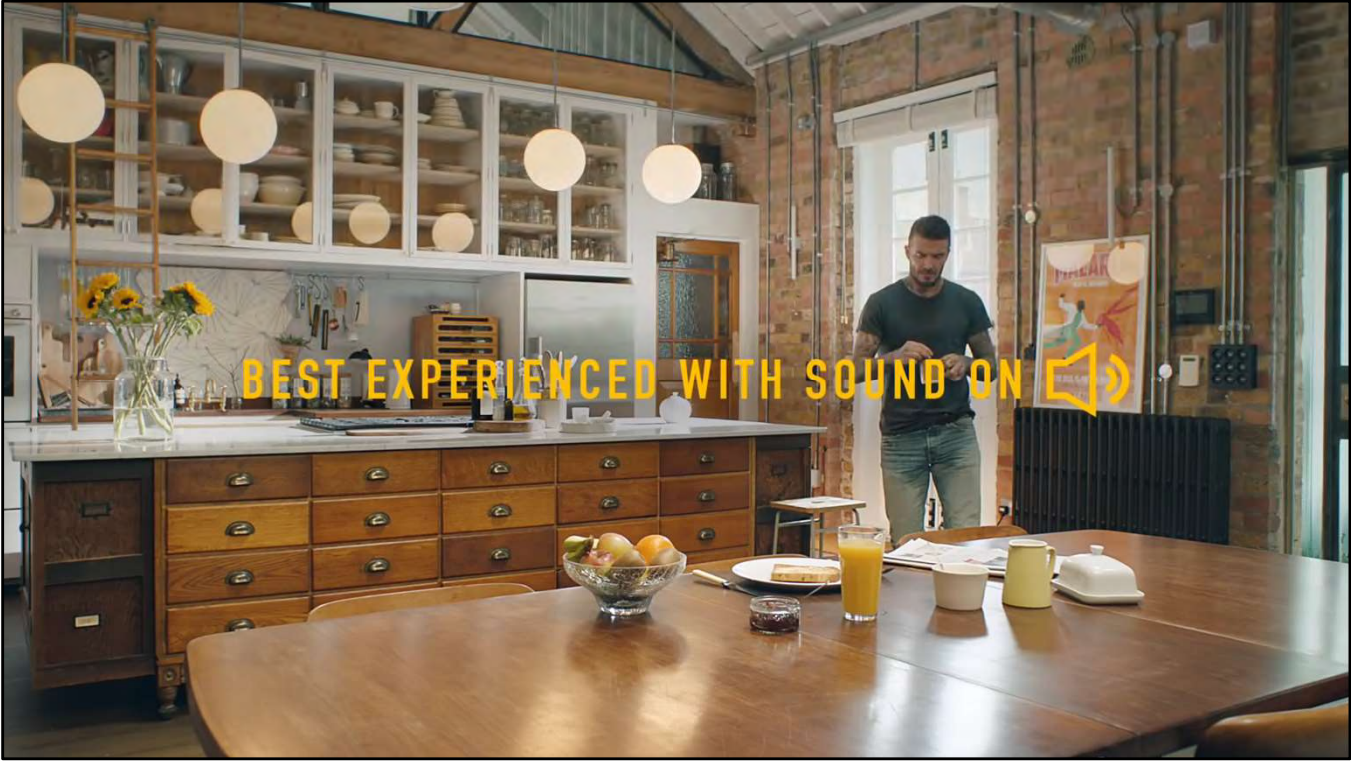
Deepfakes can simulate scientific experiments or complex and abstract concepts, making them more tangible and accessible to students. Through visualizations and interactive simulations, students can explore scientific principles and engage in virtual experiments that might otherwise be difficult to replicate in a traditional classroom setting.

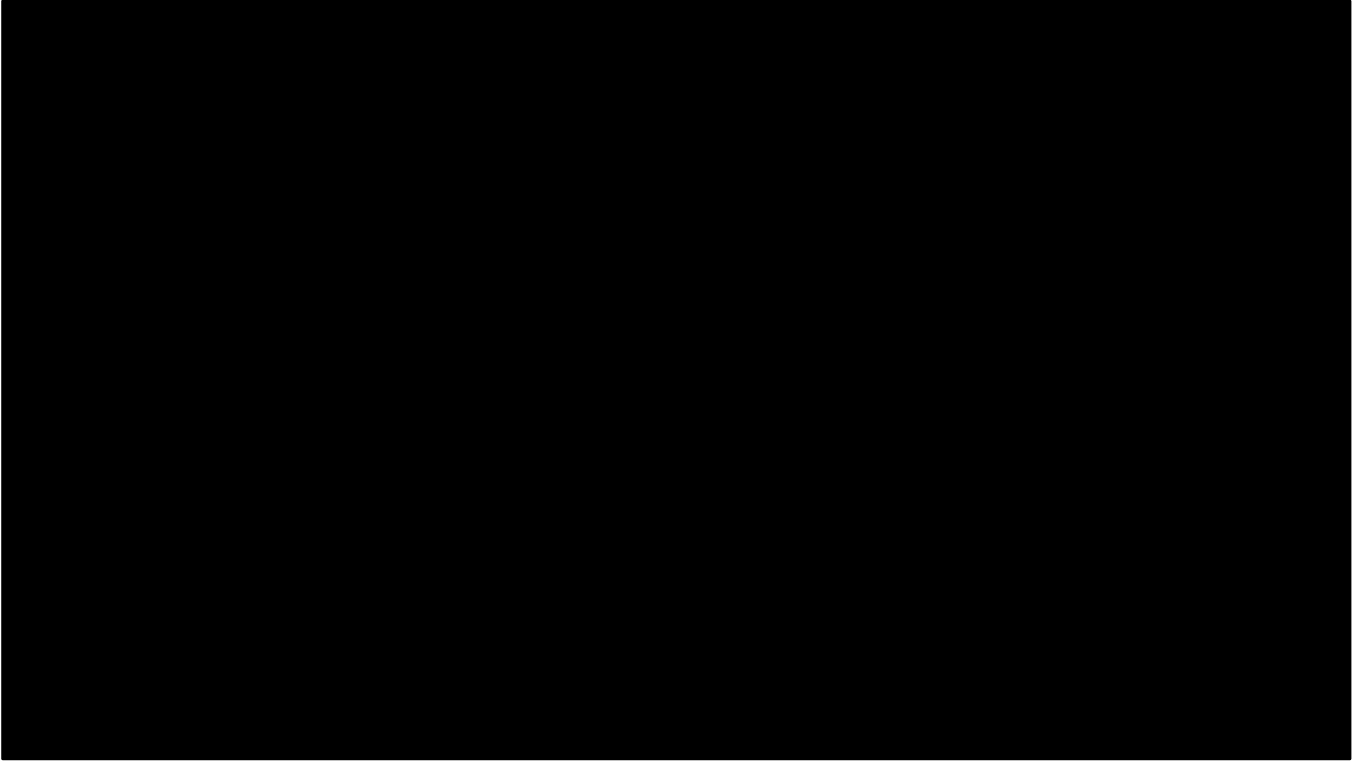
Moving out of purely educational territory, deepfakes have the potential to assist individuals with disabilities. For example, they can help people with speech impairments

to communicate more effectively by generating natural-sounding voice output based on their own facial movements or gestures.

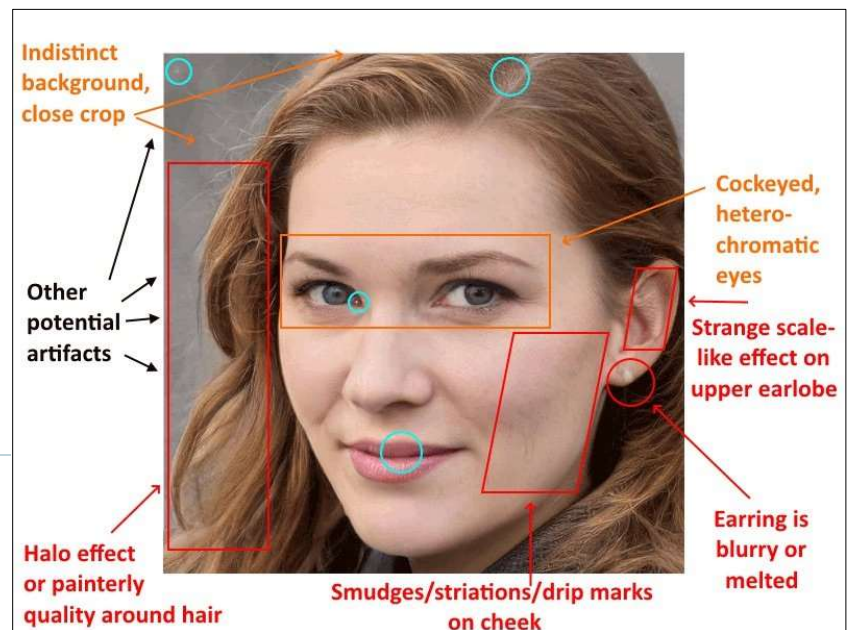
It can be employed to restore and enhance old or damaged film footage. By digitally recreating missing or deteriorated segments, films and historical recordings can be preserved for future generations.

Finally, can be used to dub movies or TV shows, enabling seamless lip-syncing in different languages. This can make content more accessible to global audiences while preserving the original actors' performances. I'll show some examples of the last two points. First, we'll see how it's used for dubbing in video featuring David Beckham speaking up to nine different languages to spread awareness about malaria and how the disease can be brought under control. And then there'll be a couple minutes from Peter Jackson's World War 1 documentary *They Shall Not Grow Old* where they restored and modernized footage, adding colour and audio where there was none...





What to look out for



With all that in mind then, it's worth looking at some of the red flags you need to look out for and how you can spot when something isn't "real". Looking at the fake LinkedIn account of "Katie Jones" I mentioned earlier, there're a few things that were identified as looking unnatural. I think the most obvious ones are on the ear where there's the strange scale effect and the blurry earring, but I'll go through a list of some of the basic things to keep on eye on...

- ~~Deepfakes don't blink~~
- Unnatural facial movements
- Inconsistent or unnatural lighting
- Blur or artifacts around edges
- Lack of eye contact
- Unusual head or body movements
- Vocal inconsistencies
- Contextual inconsistencies
- Check multiple sources



Some deep fakes are obvious

but as the technology improves it gets harder to spot them. For example, in 2018 researchers discovered that deepfake faces tend not to blink as the pictures used tend to be of people with their eyes open and therefore the algorithms don't learn about blinking. But soon as that research was published,

deepfakes appeared with blinking. So as soon as a weakness is revealed, it's fixed.

However, there are often still unnatural Facial Movements: Deepfakes often exhibit subtle or noticeable irregularities in facial expressions and movements. Unnatural blinking patterns, awkward lip movements, or misaligned facial features that don't match the rest of the face.

Inconsistent or Unnatural Lighting: Pay attention to lighting and shadows in the video. Deepfakes may have lighting inconsistencies, such as mismatched shadows or lighting angles that don't align with the surrounding environment or the person's face.

Blur or Artifacts Around Edges: Deepfake algorithms can sometimes struggle with accurately blending the manipulated face with the background. Check for blurriness, pixelation, or distortions around the edges of the face, especially when there is

movement or changes in the camera perspective.

Lack of Eye Contact or Gaze: Deepfakes may not have convincing eye contact or natural gaze patterns. The eyes may appear flat, not tracking objects or people properly, or looking slightly off in the video.

Unusual Head or Body Movements: Deepfakes can sometimes introduce unnatural head or body movements due to limitations in the algorithms. Look for strange or exaggerated movements, unusual head tilts, or body posture that doesn't align with the context of the video.


Inconsistencies in the voice: If the deepfake involves manipulated audio, listen carefully to the voice. Look for odd fluctuations, robotic tones, mismatched lip-syncing, or unnatural pauses that may indicate audio tampering.


Contextual Inconsistencies: Consider the context of the video or the source it is coming from. Does it align with what you know about the person or the situation? Look for discrepancies in the narrative, background details, or events that seem unlikely or out of character.


Check Multiple Sources: If you encounter a video that seems suspicious, try to find the same or similar content from multiple sources. Comparing different versions can help identify inconsistencies or discrepancies that may indicate a deepfake.


 Develops critical thinking skills

Verifying sources

 Cross-referencing and fact checking

 Recognising manipulation techniques

 Awareness of emotional manipulation

 Ethical implications

So going forward, I think the issue with relying on technology to identify deepfakes is that it will always be playing catch up and whilst those tools should continue to be developed, I think what really needs developing is everyone's media. Now, very unfairly and I'm not just saying this because I've got a degree in media, but media seems to be looked down upon by a lot of curriculum areas as somehow lesser which I've always found baffling as apart from breathing oxygen, it's arguably the thing people in the western world engage with the most. Whether it's watching TV, reading a newspaper, mindlessly scrolling through Instagram on your phone or whatever, we all need the tools to be able to navigate the digital world and we need to foster in the next generation of informed and responsible digital citizens. Deepfakes pose a significant challenge to media literacy, and educating people about the existence and implications of deepfakes is crucial in fostering a discerning audience and gives folks the skills to critically analyse and evaluate the credibility and authenticity of digital content they encounter. Some of the ways greater media literacy can help navigate the digital world...

Deepfakes require individuals to engage their critical thinking skills. Encouraging scepticism and questioning the information presented is vital. Teaching individuals to assess the source, context, and quality of media content helps them navigate the digital landscape effectively.

Following on from the point on the previous slide, media literacy emphasizes the

importance of verifying sources before accepting information at face value. Deepfakes often manipulate or misrepresent the source of information. Teaching individuals to verify the credibility of sources can help identify and avoid falling victim to deepfake content.

Cross-referencing information across multiple reliable sources and fact-checking claims. By encouraging individuals to seek corroboration from trusted sources, they can minimize the impact of deepfakes and misinformation.

Media literacy educates us on common manipulation techniques employed in deepfakes. Including understanding visual anomalies, audio inconsistencies, and contextual clues and so on as mentioned before.

Deepfakes can exploit emotions and elicit strong reactions. Media literacy helps individuals recognize emotional manipulation and prompts them to evaluate the content with a rational and critical mindset.

Addresses the ethical considerations surrounding deepfakes. It encourages individuals to think about the consequences of creating and sharing deepfake content, emphasizing the importance of responsible digital citizenship.

Now before I finish I'm gonna play one last video, it's only a minute long but I think it's a pretty good encapsulation of deepfakes. Have a look at what things stand out to you to show that it's not genuine. Just to warn you there is some flashing images at the end directly after the chap in the video asks "Now what do you see?" that lasts a couple of seconds so keep that in mind if you're someone who's effected by that.



We have reached the end

So that brings us to the end, I hope you're all feeling slightly more media literate than you were before, and you've got a better idea of how to navigate the era of deepfakes but also of how to engage with all digital content responsibly. It's good to have a healthy dose of scepticism and to not take everything at face value. But as William Shakespeare famously wrote...

“Beware, dear souls, for deepfakes be the jesters of deceit, masking reality with a villainous feat.”
– William Shakespeare



“Beware, dear souls, for deepfakes be the jesters of deceit, masking reality with a villainous feat.”

Thank you!