Deepfake Forensics: Exploring the Impact and Implications of Fabricated Media in Digital Forensic Investigations

Dr Áine MacDermott
BSc (Hons), PhD, PGCertHE, FHEA

School of Computer Science and Mathematics,
Liverpool John Moores University, UK.

dream plan achieve

1

Deepfake

Deepfake definition from Oxford Languages:

noun

A video of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information.

"*the committee hearing on worldwide threats cited deepfakes as a growing concern*"

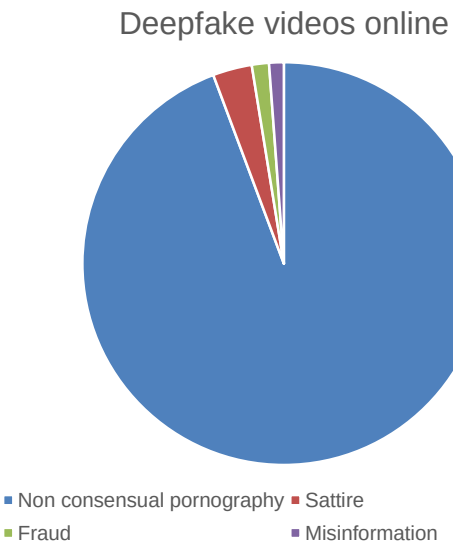
The rise in fabricated media - ‘deepfakes’

- Deepfake technology is continually evolving, becoming more sophisticated and harder to detect. There has been a concerning rise in their malicious use including non-consensual pornography and various criminal activities such as defamation, fraud, and spreading misinformation.

- As such, deepfake detection and analysis pose significant challenges for digital forensic practitioners.

- This talk explores the impact of deepfakes on criminal investigations and digital forensic practitioners, the growing influence of fabricated media, and the challenges in identifying and combating manipulated digital media.

Deepfake trends according to DeepTrace report



- Deepfake pornography is often described as image-based sexual abuse -- a term that also includes the creation and sharing of non-fabricated intimate images.
- Current laws do not go far enough to cover disturbing and abusive new behaviours born in AI and smartphone era.
- The UK government is introducing a Crime and Policing Bill aiming to crackdown on explicit deepfakes.

Deepfake and online fraud

Deepfake technology is a rapidly advancing field that has the potential to revolutionise entertainment, advertising and even politics. However, it also poses a serious threat to online security, as deepfakes can be used for malicious purposes like online fraud. In this blog, we explore how deepfake technology is making this criminal behaviour more difficult to detect and what individuals and businesses can do to protect themselves.

The rise of deepfake porn and associated tools

Alongside the rise of headline-grabbing tools such as ChatGPT, there are a plethora of other easily accessible apps and websites that allow users without technological expertise to generate fake pornographic images that transpose real women's faces onto explicit pictures. Simple, fast and cheap, the quickly emerging technology provides a terrifying new tool for abusers to target women and girls, with 96 per cent of deep fake images on the web estimated to be pornography and the vast majority featuring female subjects.

> **Users can create content featuring real-life women with near-total impunity**

The potential harms are wide-ranging, from the use of such images to coerce and blackmail women, to the long-lasting psychological, emotional and reputational damage on victims. Already there have been reports of women losing their jobs as a result of deepfake pornographic images made of them without their participation or consent. Like other forms of online and image-based abuse, there is a misguided tendency – often from those with no direct personal experience of the issue – to dismiss this as a trivial issue that victims should simply be encouraged to ignore, but this entirely fails to acknowledge the enormous impact on many aspects of victims' lives, as well as the significant damage to their mental health.

Nonconsensual media

Criminal reforms target 'deepfake' and nonconsensual pornographic imagery

'Downblousing,' 'upskirting' and sharing 'deepfake' pornography without consent could lead to jail sentences of up to three years

Campaigners say proposed reforms stop short of what is needed. Photograph: Mint Images - Tim Robbins/Getty Images/Mint Images RF

Secretly videoing or taking photographs of people under their clothes or sharing "deepfake" pornography without consent could lead to prison sentences of up to three years, under recommendations by the Law Commission of England and Wales.

"Sharing intimate images of a person without their consent can be incredibly distressing and harmful for victims, with the experience often scarring them for life," said Prof Penney Lewis, the law commissioner for criminal law.

Lewis said these offences were currently dealt with under a "patchwork" of criminal offences that had not kept pace with technology.

Gaps in the law enable perpetrators to evade prosecution.

ARTIFICIAL INTELLIGENCE · **Published** June 12, 2023 2:00am EDT

AI ‘deepfakes’ of innocent images fuel spike in sextortion scams, FBI warns

The FBI said AI has played a major role in reports of sextortion soaring

'ox News

In the US, the number of nationally reported sextortion cases increased 322% between February 2022 and February 2023.

Innocent pictures or videos uploaded to social media or sent in messages can be twisted into sexually explicit, AI-generated images that are "true-to-life" and nearly impossible to discern, the FBI said.

Predators, who are typically in another country, weaponize the doctored, AI images against juveniles to coerce them.

Sharing deepfake pornography could soon be illegal in America

The rise of AI-generated content exploded due to accessibility and ease of use.

By Emmanuelle Saliba
Video by Jessie DiMartino
June 15, 2023, 11:09 AM

Looking at how nonconsensual deepfake porn targets women….

Experts say a large majority of AI-generated videos are pornographic, nonconsensual and include more women than men.

It's a global effort…..demystifying deepfakes!

- The continuingly improving capabilities of deep learning, GANs and commercial deepfake tools alongside the threat posed by malicious actors to commit criminal acts has resulted in global efforts from researchers, corporations, governments, and law enforcement to work towards developing robust deepfake detection models that could assist in identifying whether multimedia is real or fake.

- Initiatives such as Microsoft's Project Origin (Microsoft, 2022) or the wider global partnership of Trusted News Initiative (BBC, 2022) are a few examples of the various organisations that are trying to tackle this growing issue of misinformation.

Key features and techniques of deepfakes

- Head puppetry: This technique uses a video of someone moving their head or making an expression and then applying a synthetic face of the target onto the original video to make it appear that this person is making those expressions or movements.

- Lip syncing and expression swap: This usually focuses solely on manipulating an existing video of a subject and is used in combination of synthetic audio to give the illusion that a subject is saying something or changing their expression to make it seem they are happy or angry.

- Face swapping: This technique involves swapping the face of a subject within an image or video with another to make seem they are present or involved.

- Face synthesis: This generates an entirely synthetic face for creating deepfake images and has been used by malicious actors for identity fraud or document fraud

Key features and techniques of deepfakes

- Audio-based methods: Specifically, with audio they focus on three different techniques.
  - Imitation-based deepfake methods aim to transform existing audio to make it appear that someone else is speaking the words.
  - Synthetic based methods which are what is used for text-to-speech devices that aim to read plain-text and produce a voice that sounds human.
  - Replay-based methods which aims to record the voice of the target to imitate someone's voice (using segments of real audio with injected synthetic audio).

Deepfake and law enforcement challenges

- Deepfake forensic research is increasingly gaining momentum after criminal applications of deepfake creation tools and techniques occur.
- This emerging form of synthetic media being applied for malicious means has presented several challenges to researchers and law enforcement.
- One key issue is budget concerns and the available tools.
- Time constraints associated with identification and analysis of deepfakes.

Deepfake crime

- The investigation of deepfake crimes will require an ad-hoc approach. Law enforcement personnel can look for other methods of corroborating video or refuting its authenticity such as cross-checking videos with independent sources, collecting phone records including data and GPS, and interviewing other people seen in the video. Developing methods to verify chain of custody and authenticity will be critical to maintain public trust in video evidence.

- Many in policing are still unaware of how sophisticated and easy the creation of deepfake media has become.

- Lack of deepfake laws and repercussions for crimes*

- Rise in fully photorealistic child pornography

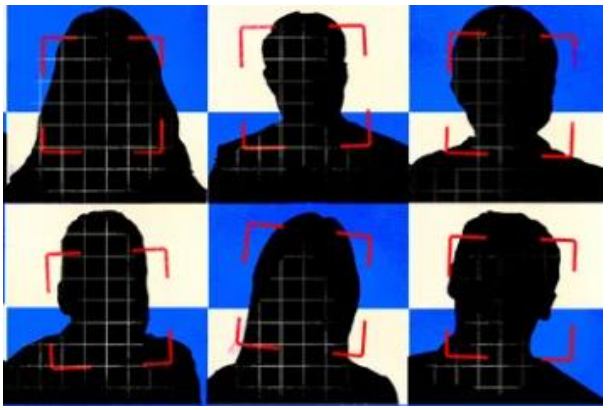
Adversarial nature of detection and anti-forensics techniques

- Malicious actors can add intentional permutations to their deepfakes to trick detectors.

- These types of anti-forensic techniques can be broken down into two types of attack: ‘causative’ and ‘exploratory’ attacks.
  - An example of a causative attack involves poisoning training data with bad samples.
  - An exploratory attack would involve a malicious actor probing the classifier to reverse engineer the model and possibly hand craft data to avoid or trick detectors.

Model generalisation

- This refers to a model's ability to detect deepfakes reliably regardless of what techniques have been used to create the deepfake.
- Many deepfake models perform well on their trained data and specific datasets but when deployed into real-world scenarios their performance is inconsistent.
- In the case of deepfake detection, networks mostly focus one form of multimedia (images, video, or audio) and may focus in on specific features or signs for detection.
- Ideally it would be better to try and create detectors that are more 'generalised' to detect various types of deepfakes created with different methods or features.



Indicators for detecting deepfakes

- Unnatural backgrounds: Deepfake media may have backgrounds that are not consistent with the original environment. Double encoding analysis can identify if a video has been compressed more than once.
- Inconsistent facial expressions: Deepfake videos often have subtle inconsistencies in the facial expressions of the person being manipulated.
- Unnatural movements: Deepfake videos may have movements that are unnatural or not consistent.

Indicators for detecting deepfakes

- Audio inconsistencies: The audio in a deepfake video may not match the lip movements of the person being manipulated or may e[illegible] the sound quality.

- Image search: Sometimes deepfake detectors report [illegible] is 'real'; however, 'real' in these cases may suggest that [illegible] fact a stock image or reference photo, which has beer[illegible] deepfake.

- Inconsi[illegible] deepfa[illegible]

Format	PNG	Format is not a standard one
Format Description	Portable Network Graphics	
Image Encoded Size	1162 x 1162	Odd width (1162 is not multi Odd height (1162 is not mult
Image Displayed Size	1162 x 1162	
Image Normalized Size	1162 x 1162	
Aspect Ratio	1.00	
Number of Channels	4	Channels count different fro
BPP	32	BPP (32 different 24)
Thumbnail Size		Thumbnail is missing

Current deepfake tools

- Look for reference images, e.g., Flickr, reverse Google Image search, stock photos, etc. They use reference image and current 'suspect' file analysis comparison
- Some determine whether the media is 'real' – 'real' means non deepfake but it could be manipulated in other ways not detected by tool
- Size of image can affect accuracy: You may need to rescale and adjust image properties to improve analysis metrics
  - Amped Software (and other tools) can create a collage of collected media, rescaling them to get more accurate results.
  - Batch processing and analysing a collection of images means you can pre-process filters and identify '*things to look out for'.*

Deepfake Experiments - DFDetect

- Aim to determine whether the media is ‘real’ – ‘real’ means non deepfake but it could be manipulated in other ways not detected by available tools! We created 100 deepfake test data images (50 real and 50 deepfake) from “ThisPersonDoesNotExist” and “ThisPersonExists”.

- We ran the images through a popular online deepfake detection tool ‘DFDetect’ and then through a premium tool ‘Amped Authenticate’. DFDetect generates a percentage score on how likely the image is to be authentic. For DFDetect, accuracy on identifying deepfakes was low but high for real images.

Table 1: DFDetect Results

No.	DFDetect (Deepfake)		
1	**100%**	11	100%
2	100%	12	100%
3	**1.23% fake**	13	99.23%
4	100%	14	100%
5	**100%**	15	95.20%
6	100%	16	100%
7	**70% fake**	17	99.55%
8	100%	18	100%
9	100%	19	99.99%
10	100%	**20**	**51.31%**

Figure 1: DFDetect vs Amped results - image numbers (left to right) 1, 3, 5, 7, 20.

Deepfake Experiments – Amped Authenticate GAN

- For Amped Authenticate, there was a good improvement on accuracy but many deepfake samples scored 50-60% or were deemed ‘uncertain GAN’. We tested on a range of ages, genders, and angles and are expanding to include more complex photos.

Table 2: Amped Authenticate GAN Deepfake Detection Results

No.	Amped Authenticate (Deepfake)		
1	77% Not GAN	11	92% Not GAN
2	**73% Uncertain GAN**	12	79% GAN
3	83% GAN	13	100% GAN
4	**99% Not GAN**	14	**95% Not GAN**
5	71% Uncertain GAN	15	99% GAN
6	**62% Uncertain GAN**	16	**62% Uncertain GAN**
7	79% GAN	17	81% GAN
8	**87% Not GAN**	18	**67% Uncertain GAN**
9	**66% Uncertain GAN**	19	96% GAN
10	76% GAN	20	97% GAN

Figure 2: Uncertain GAN Amped results

Conclusion

- Deepfake use in crime is on the rise and this is a growing challenges for police and digital forensic units. Deepfakes can obscure the truth or introduce doubt about the authenticity of critical evidence.
- The time spent analysing a suspect deepfake to confirm the validity of the media in more time consuming than non deepfake and the availability/accuracy of tools can vary.
- Also, attackers may exploit online training datasets, corrupting AI-driven forensic tools with fake media. Or by manipulating metadata or embedding fakes within legitimate datasets, attackers can complicate detection and attribution.

Thank you for listening. Any questions?

@ainemacd

a.m.macdermott@ljmu.ac.uk

I invite you to take part in a research survey:

dream plan achieve