Deepfake Evidence in Criminal Proceedings

Procedural hurdles and forensic challanges

Clementina Salvi
Doctoral researcher at Queen Mary University of London
Research Assistant in EU Criminal Law and Digital Justice at the University of Liverpool

FACULTY OF LAW, ECONOMICS AND FINANCE
DEPARTMENT OF LAW
Fonds National de la Recherche Luxembourg

Supported by the Luxembourg National Research Fund (FNR) C20/SC/14755738)

Outline

Origin and understanding

Deepfakes in criminal proceedings

key legal and procedural challenges

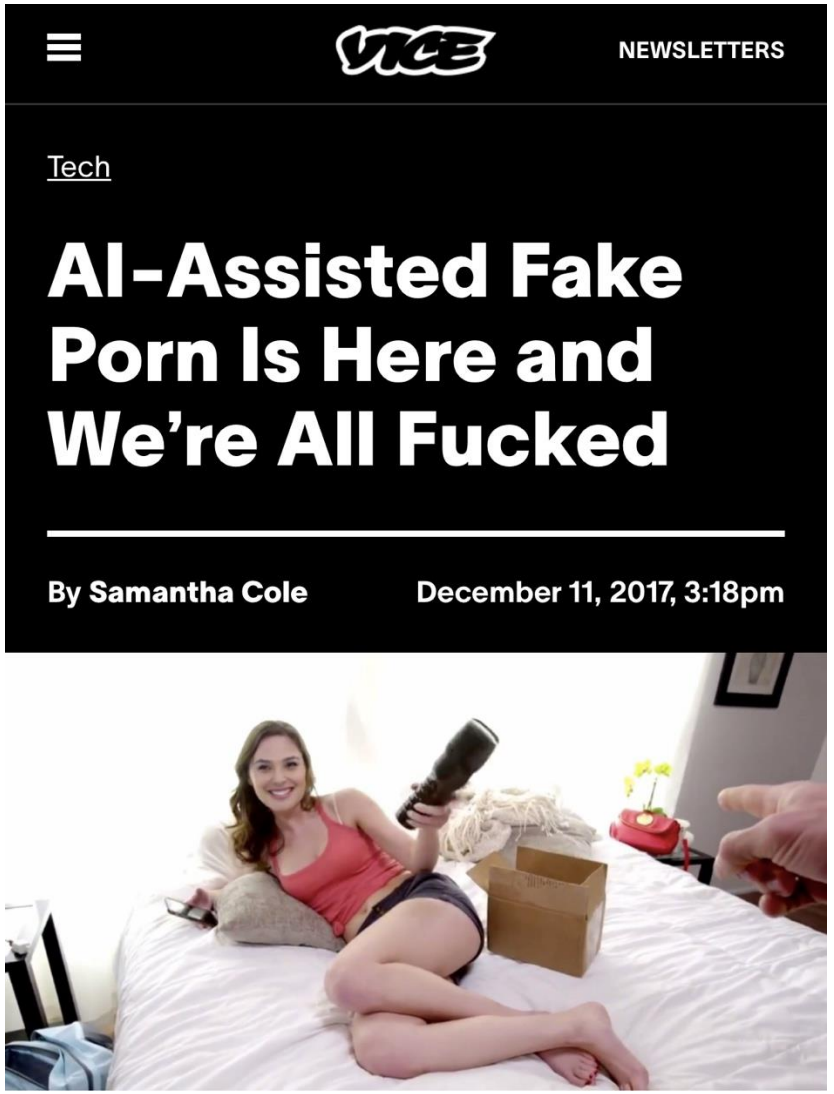
Proposed solutions

The emergence of Deepfakes

‘Deep fakes’ – a combination of ‘deep learning’ and ‘fake’ – 2017: Reddit user under the nickname *deep fakes* posted a number of videos of famous actresses and singers with their faces superimposed on the bodies of women in pornographic films

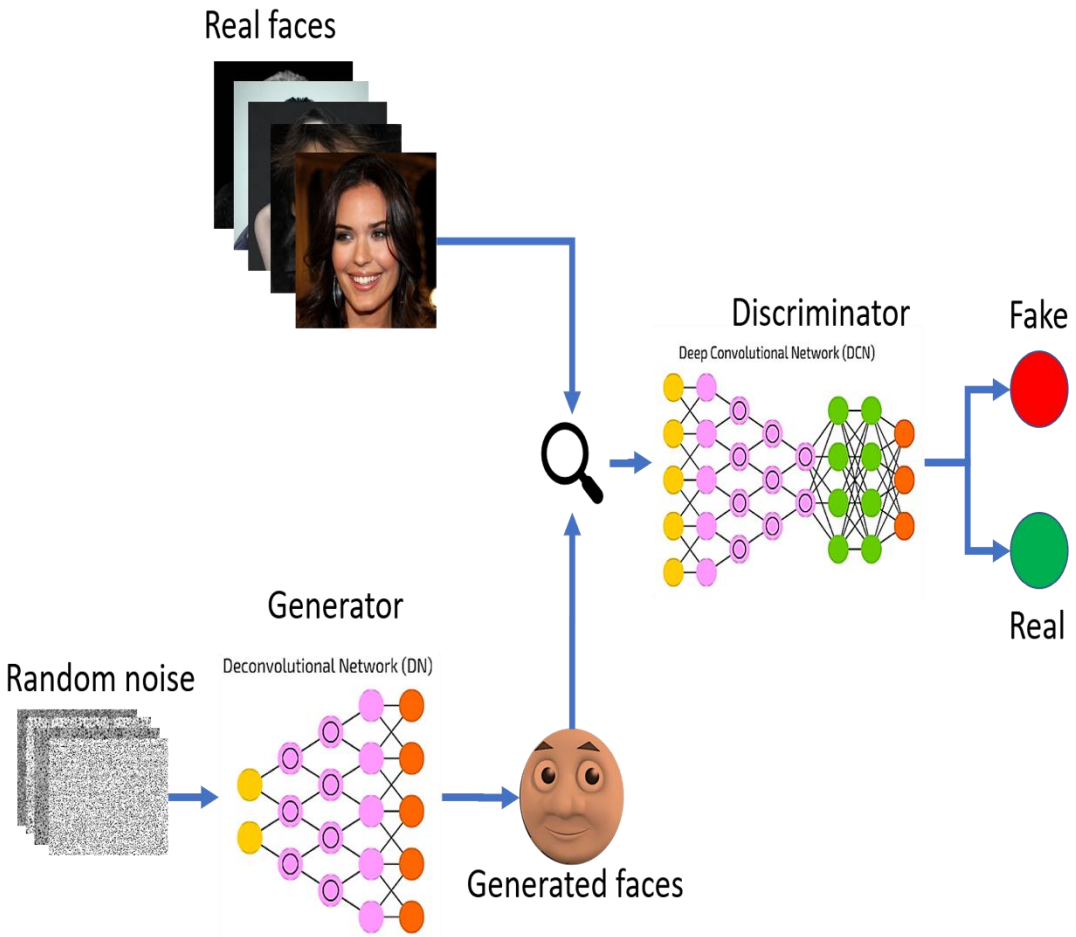
Deep learning is a branch of Machine Learning: AI systems that have the capacity to respond to, as well as to create language, images and sounds

Use of AI to ***manipulate, alterate*** or ***generate synthetic media*** that can be difficult for humans or even technological solutions to distinguish from authentic ones



The technology behind

- **Generative Adversarial Network** (GAN): two different ANN algorithms compete against each other. One creates the fake content (the forger), while the other tests whether the content is real or produced by a computer (the discriminator).
- **Diffusion models:** used to create Picture to Picture transformations that allow for smooth transition between image distributions
- **Autoencoders :** type of neural network that are trained to efficiently encode input data, capturing essential features while discarding irrelevant information. They excel at face swaps using its ability of retaining image quality while seamlessly replacing facial features in videos deepfakes.
- .... And others!



Added negative value

- **Multi-modal and highly-realistic**
- **Accessible and easy to produce**
- **State of the art is improving rapidly**
- **Detection efforts and the cat-and-mouse game**
- **Potential harm and public perception**

Expert believe that 90% of the content available online will be synthetic (Interpol, 2024)

ThisPersonDoesNotExist.com

The rise of Generative AI

- GenAI hype 2022: AI technology for digital content generation (Chat-GPT; Stable Diffusion; Dall-e)
- AI that outputs entirely new pieces of synthetic media
- Beyond initial prompting, the creative process is fully automated - no longer does one need to swap faces into pre-existing videos (Citron, 2022)
- Use of AI text-to-image generators: interactive and conversational realistic-looking AI-avatars

The New York Times

Definitional issues

- Are deepfakes just one of many “synthetic media”?
- Does a deepfake have to be hyper-realistic to be defined as such? It can be hyper-realistic but not generated through AI or *vice-versa* (dichotomy deepfakes - cheap fakes and deepfakes - shallow fake)? The understanding is also strongly rooted in social consciousness, where deep fakes are directly associated with impersonating other people → existing people
- Lack of precise or shared definition for deepfakes - rapidly evolving technology that underlies them
- no *consensus* on the media covered by the word, if it encompasses only videos and images, later on also audios, and what about texts? Both academic and scientific descriptions have so far been focusing on audio and visual images (video and pictures), leaving outside other kinds of media, such as AI-generated texts (for the inclusion Farid, 2022)

AI ACT

Article 3(60) describes a “deep fake” as an AI-generated or manipulated image, audio or video content ***that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful***

https://artificialintelligenceact.eu/

Scenarios involving deepfakes in criminal proceedings

- as evidence of deepfake related crimes (King, Floridi and others 2020)
- as tampered evidence *in* criminal trials to both frame or exonerate individuals (Alexandrou and Maras 2018, Chesney and Citron 2020, Grimm 2021, Delfino 2023)
- to enable and enhance images; montage generation (eg. translate eyewitness descriptions into visual representations of potential suspects or for creating virtual simulation of crime environment); LEAs undercover operations (Interpol Report 2024; van der Sloot 2024)

DETECTION TOOLS

forensic assessment is crucial

Deepfake detection and authentication tools

Detection tools i) to *authenticate* or assess reliability of digital evidence in general, as well as ii) *source* of evidence for criminal fact-finding, specifically for crimes committed or facilitated through deep-fakes (**output as evidence**)

- No naked eyes (Wu and Liu, 2019; Verdoliva, 2020; Farid, 2022)
- Limits of human expert analysis (Palmiotto, 2023)
- AI-based detection tools: 'fight AI with AI' automatically (Giudice, Guarnero, Battiato, 2019)

Examples (Interpol 2024):
Deep Learning Models
File Structure Analysis
Biological Signals
Statistical Analysis at a Pixel Level
Geometrical and Behavioral Analysis

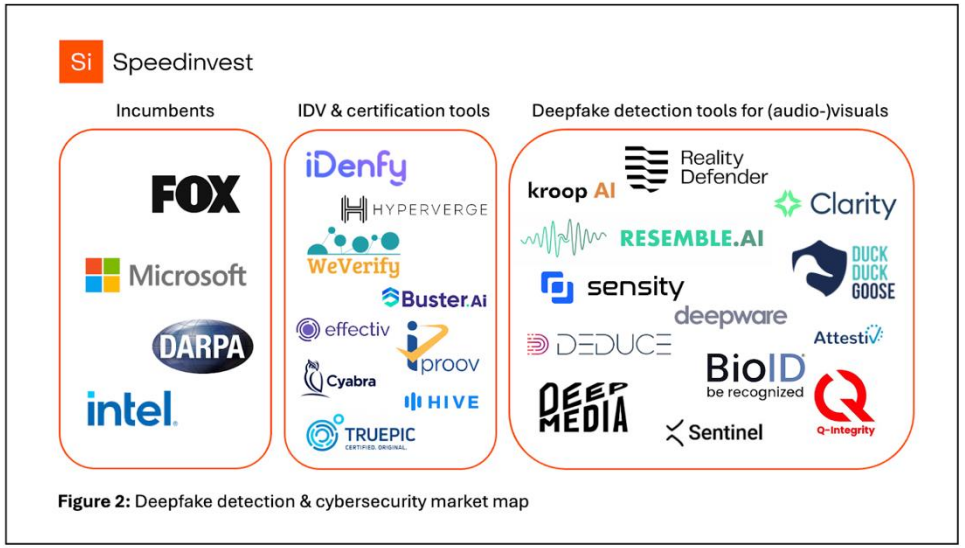


Figure 2: Deepfake detection & cybersecurity market map

Authenticity

- deepfakes can undermine the “presumption of authenticity” of visual media, as they are often indistinguishable from real media even to forensic experts
- Importance of digital evidence in criminal trials (user-generated videos, screenshots of WhatsApp conversation and so on)
- Low conviction rate for cases of sexual violence often justified by the so-called ‘he-said-she-said’ nature of these cases, the increasing presence of digital evidence has begun to challenge this justification (Dodge 2017; Powell 2015)
- Black Lives Matter (BLM) movement, it has been highlighted the importance of citizens capturing to ensure that photos and videos taken on smartphones and other devices are stored in ways that establish their authenticity in courtrooms (Zarmsky 2020)
- Are current rules robust enough to survive the authenticity challenge?

Detection, procedural fairness and access

- Even if detection systems are flourishing, there are yet no shared standards or regulations that ensure detective tools are “enough” accurate. This is extremely complex considering that every media is subject to a certain degree of manipulation, integrated in digital cameras, software, filters and so on
- Who determines the accuracy and reliability of AI detection tools? (Palmiotto 2023): the output of AI-detection tools/algorithms can be used as substantive evidence in criminal prosecution in deepfake related crimes
- Ensuring fair access *to* detection tools is challenging*:* prosecutors may have resources to access advanced deepfake detection tools, but defense teams, especially in low-resource cases, may not. This disparity raises issues around “equality of arms” (Palmiotto 2023)
- Risks associated to AI anti-deep-fakes tools which *automatically* verify the authenticity of the media. So-called ‘black box’ and obscure algorithms issues raise here. In certain cases, none of the experts may be capable of explaining the outputs of algorithms

Undermining the credibility of digital evidence "It wasn't me"

- Suspects and convicted could claim that incriminating video, image or audio evidence is fake, undermining the credibility of potentially legitimate evidence ("**deefpake defence**")
- The "**liar's dividend**" issue (Chesney and Citron 2019). The consequence might not only be that untrue events are perceived to be real, but that real events might also be perceived to be unreal. To some this is effect can undermine trust on visual media on a large scale and bring a "credibility crisis" in legal contexts
- In the ***UK case People v. Foreman, 2020 IL App (2d District) 180178***, a case of first-degree murder and residential burglary, the defendant argued that his voice was cloned. He argued that 'in the age of so-called deep-fake videos and easily-manipulated audio recordings, improperly authenticated recorded communications should be inherently suspect'. The Court rejected the defendant's argument 'that recent technological advancements render all recordings suspect, because they can be easily manipulated. In the absence of any evidence of tampering or other such manipulation in this case, there are no foundational issues with the recordings'. However, the case shows how this argument can potentially impact on the length of the proceeding and current interpretation may soon change.

Technological safeguards: watermarking, transparency duty and safety by design

- The use of watermarking and safety by design measures (such as fingerprints) that will assure authenticity – support of AI companies—Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI
- marker that indicates the file's origin or authenticity and can reveal tampering if the media is altered

Challenges and Limitations

- While watermarking shows promise, it is not a standalone solution
- **Wide adoption and compliance**: effective watermarking would require widespread implementation and compliance across all media creation platforms
- **Circumvention Risks**
- **Legal Standards**: Courts may need standardized guidelines for interpreting watermark evidence in criminal proceedings

Transparency duty: the AI ACT

Art. 50– Transparency obligations for providers and deployers

(2) Providers of AI systems, including general-purpose AI systems, **generating synthetic audio, image, video or text content**, shall ensure that the outputs of the AI system are marked in a **machine-readable** format and detectable as artificially generated or manipulated. Providers shall ensure their **technical solutions** are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant **technical standards**. (…)

(4) Deployers of an AI system that **generates or manipulates image, audio or video content constituting a deep fake**, shall disclose that the content has been artificially generated or manipulated. **This obligation shall not apply where the use is authorised by law to detect, prevent, investigate or prosecute criminal offence**. Where the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.

Deployers of an AI system that **generates or manipulates text** which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated. (…)

New evidentiary rules

- To ensure procedural fairness, courts may need to adopt evidentiary standards to authenticate digital evidence and verify that evidence is genuine as well as adapt legal standards for admissibility domestically
- In the US debate is vivid also includes several proposed rules and amendments to the Federal Rules of Evidence (the general principle is already that evidence must be authentic and reliable) specifically aimed at addressing the challenges of deepfakes (LaMonaga 2020; Delfino 2023; Grimm and Grossman 2024).

At the April 2024 meeting of the Advisory Committee on Evidence Rules, Judge Paul Grimm (Ret.) and Dr. Maura Grossman made a presentation about the evidentiary problems caused by deepfakes and proposed a new Fed. R. Evid. 901(c):

Potentially Fabricated or Altered Electronic Evidence.

If a party challenging the authenticity of computer-generated or other electronic evidence demonstrates to the court that it is more likely than not either fabricated, or altered in whole or in part, the evidence is admissible only if the proponent demonstrates that its probative value outweighs its prejudicial effect on the party challenging the evidence

ELI Proposal

Eli proposal art. 7 on admissibility of electronic evidence: all digital evidence in court should be assessed on authenticity

Paragraph 3 requires Member States to ensure, through national rules, that **electronic evidence is *not used in criminal proceedings unless there is sufficient evidence that it is not the result of manipulation or forgery***. Given that instances of digital image processing (eg. deepfakes) and other kinds of data manipulation are difficult to trace, **an unchecked use of electronic evidence shall no longer be permissible**; instead, **it shall be specifically checked whether the electronic evidence is *not* the result of such manipulations.**

In order to check whether the requirements of paragraphs 1 to 3 are met, and therefore whether the electronic evidence has not been altered in terms of content and scope between the time of obtaining and using it, and whether it is not the result of manipulation and forgery, ***it is essential to have access to the expertise of IT experts****. According to paragraph 4, Member States are therefore obliged to allow the involvement of IT experts at the request of the suspect or accused person. However, Member States do not necessarily have to bear the costs for these IT experts, although this is recommended as long as it contributes to* ***the fairness of the proceedings and the equality of arms.***

Developing forensic standards and trustworthy AI based detection

- Development of shared good practice and standards at forensic level to guarantee levels of accuracy
- AI detection tools treated as “expert scientific evidence” even when used just to authenticate evidence (Palmiotto, 2023) – i) ***transparency*** and ii) ***explainability*** and iii) ***scientific validity*** of the method

- Ensure ***fair access*** to detection tools

Enhanced judicial training

- Judges and legal professionals may require **training** in media forensics and AI detection to accurately interpret complex digital evidence
- Such training would enable the judiciary to more effectively evaluate the reliability of deepfake evidence
- *«Do your due diligence»* understanding signs of deepfakes, verify provenance before offering media as evidence, and be prepared if authenticity is questioned. «If a “smoking gun” video seems too good to be true, it probably is» (Pfefferkorn, 2021)

Thank you for your attention

Clementina Salvi
clesalvi@liverpool.ac.uk
c.salvi@qmul.ac.uk

UNIVERSITY OF LIVERPOOL

Queen Mary University of London