Sandoval *et al. Crime Science* (2024) 13:41
https://doi.org/10.1186/s40163-024-00239-1

Crime Science

SYSTEMATIC REVIEW

Open Access

Threat of deepfakes to the criminal justice system: a systematic review

Maria-Paz Sandoval[1*], Maria de Almeida Vau[1], John Solaas[1] and Luano Rodrigues[1]

Abstract

Background This systematic review explores the impact of deepfakes on the criminal justice system. Deepfakes, a sophisticated form of AI-generated synthetic media, have raised concerns due to their potential to compromise the integrity of evidence and judicial processes. The review aims to assess the extent of this threat, guided by a research question: (1) What threats do deepfakes pose to the criminal justice system?

Methods The review was conducted using databases such as Web of Science, ProQuest, Scopus, and Google Scholar, focusing on publications from 2021 to 2022. Search terms were optimised for sensitivity and specificity, and articles were chosen based on criteria including relevance to deepfake threats and deepfake detection research. The methodology included rigorous screening processes using tools like Zotero and Rayyan.ai, with an emphasis on inter-rater reliability to ensure objective selection of studies.

Results The search initially identified 1355 articles, with 1200 articles screened for eligibility after duplicates were removed. For the threat of deepfakes to the criminal justice system, 110 studies were selected for full-text review, and 44 were included in the final analysis. Key findings include identification of primary crime categories linked to deepfakes, such as pornography, fraud, and information manipulation, alongside challenges like trust erosion in institutions and evidence falsification issues.

Conclusions Deepfakes significantly threaten the criminal justice system, highlighting the necessity for advanced detection methods. These findings underscore the importance of continued research and development in deepfake detection technologies and strategies for legal safeguards and broader implications on policy, national security, and democratic processes.

Keywords Deepfakes, Criminal justice system, Emerging crimes, Technological threats

Background

Deepfakes, a term that surfaced in 2017, represent a relatively new phenomenon in the realm of synthetic media. These are produced using machine learning techniques to create altered or entirely synthetic audio, images, videos, textual content, or real-time/live formats. Initially confined to specific tech circles, the concept of deepfakes has gradually been subsumed under the broader umbrella of AI-generated synthetic media. An illustrative example of this technology is facing swapping, where one person's face is replaced with another's in a video or image. The production of deepfakes primarily utilizes deep learning methodologies, a subset of machine learning, to train generative neural network architectures like GANs (generative adversarial networks) and Autoencoders, with GANs first being introduced in 2014 (Godfellow, et al., 2014).

In recent years, the field of AI-generated synthetic media has experienced significant advancements, marked by the emergence of technologies such as DALL-E and

*Correspondence:
Maria-Paz Sandoval
Maria.sandoval@ucl.ac.uk
[1] Department of Computer Science, University College London, London, WC1E 6BT, Gower Street, United Kingdom

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Stable Diffusion. These innovations, harnessed by sophisticated algorithms, have expanded the accessibility of synthetic media creation to a broader audience, moving beyond the confines of specialized research. For example, DALL-E employs a variation of the GANs used in deepfakes, but it is designed to generate original, high-quality images from textual descriptions. In parallel, Stable Diffusion signifies a significant advancement in generating detailed and diverse synthetic media. These developments serve to highlight the rapid progress in generative AI and the need to comprehend and address their broader implications. Given that deepfakes are now part of this expansive category of synthetic media, it becomes imperative to scrutinize their impact, especially in sensitive areas such as the criminal justice system. This systematic review, therefore, positions deepfakes within this evolving technological context, examining their potential threats and the ongoing developments in detection and mitigation techniques.

Deepfakes have notably risen to prominence over the past 5–6 years. While some deepfake applications are benign, others, like the notorious deepfake of President Volodymyr Zelensky, which emerged at the onset of the Russian invasion, showcase their potential for malicious use (Deepfake presidents used in Russia-Ukraine war - BBC News [Internet]., 2022). From a computer science perspective, substantial research has been dedicated to both the creation (Nguyen, 2022) and detection (Pianese et al., 2022) of deepfakes. However, there is a paucity of literature specifically addressing the threat posed by deepfakes to the criminal justice system, a system fundamentally concerned with the apprehension, prosecution, defence, sentencing, and punishment of criminal offenders.

This systematic review aims to analyze the intersection between deepfakes and the criminal justice system and assess their impact on the latter. With these considerations in mind, this paper seeks to provide a comprehensive review of the existing body of work concerning the threat of deepfakes to the criminal justice system. The objective will be pursued through the following research question:

RQ_1: What threats do deepfakes pose to the criminal justice system?

Methods: literature search/method[1]

Search terms

The following databases were used to conduct a search in December 2022: Web of Science, ProQuest, Scopus, and Google Scholar. We limited the search to papers published in 2021–2022 to examine the most recent research on deepfake detection and the consequences of deepfakes to the criminal justice system. Search terms were piloted to result in a balance between sensitivity and specificity, where an academic librarian was consulted to validate the search terms used. The search terms that resulted from this were:

- "Deepfake*" OR "Deep fake*" OR "Deep-fake*"
- "Court*" OR "Justice system" OR "Proof*" OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim*" OR "trial"

These keywords were used to search the article title, abstract, and index subject heading.

Inclusion/exclusion criteria

The systematic review seeks articles that discuss the threat of deepfakes to the criminal justice system, threats to reliability in court, or threats relevant to the investigation and prosecution of crime. We included articles that used a study design that were either RCTs, qualitative, quantitative, systematic review, meta-analysis, case report, or legal documents.

In terms of the type of paper we chose to limit it to the following types:

- Peer-reviewed papers,
- Journal articles,
- Academic papers,
- Policy/white papers,
- Industry reports,
- Systematic reviews,
- Meta-analysis, and
- Theses.

We excluded papers published in magazines or newspapers, those without access, and master's dissertations to ensure the credibility and relevance of our sources. Academic journals typically undergo peer review, providing a higher level of academic rigour compared to magazines or newspapers. Excluding papers without access avoids incomplete or unverifiable data. Finally, master's dissertations often lack the peer-review process, and the depth required for inclusion in high-level academic research. This approach ensures that our review is based on reliable and accessible scholarly work. We also employed exclusion criteria proposed by Edanz-Learning-Team (2022) and Meline (2006), which include:

- Issues with methodological quality
- Review articles with no original data

[1] See appendix A for Systematic Review Protocol.

- Works which are not relevant to the research question and outcomes
- Sources outside the defined time scope (2021–2022)
- Sources in languages other than English

Search strategy for identification of studies

Searches were carried out using the software Publish or Perish and the following databases were searched:

- Scopus
- Web of Science
- ProQuest
- Google Scholar

Filtering stages

After the initial search, duplicates were removed using Zotero software. Subsequently, articles were imported into Rayyan.ai, a screening tool for systematic reviews, to review the articles based on basic inclusion and exclusion criteria. After removing duplicates and excluding irrelevant articles, the remaining articles were ready for further analysis.

Inter-rater reliability (IRR)

To ensure an adequate inter-rater reliability and fairness in the selection of studies, the original coding results were checked by all four authors of this paper. At the onset of this review, we established a set of inclusion and exclusion criteria, which informed subsequent article selection decisions. An initial screening was conducted by two of the authors, with any arising conflicts resolved by a third team member. In the second and final stage of article selection, we divided into two independent teams and then compared results, resolving any conflicts amongst all four coders, based on a majority agreement. In cases where no majority was achieved, conflicts were resolved by revisiting and elaborating on the main aims and objectives of this review, realigning the discussion with the initial inclusion and exclusion criteria. Upon comparison of the results, we registered an inter-rater reliability score of 0.91, reflecting a good agreement rate between the coders (Bajpai et al., 2015 Mar; Fink, 2010; Reliability & in Systematic Review Methodology: Exploring Variation in Coder Decision-Making-, 2021).

During the initial screening phase, of the 1200 articles evaluated for eligibility, two authors independently coded a significant subset (After removing 155 duplicates). Any discrepancies were resolved with the intervention of another team member. For the 110 full-text articles assessed later, the team was divided into two independent groups to ensure objectivity in the final selections (44 articles). This rigorous process yielded a high inter-rater reliability score, confirming the consistency of our article selection methodology.

As for the analytical approach of the included studies, we utilised thematic analysis. This method allowed us to identify, analyse, and report on patterns within the data, facilitating the synthesis of findings that are crucial to understanding the implications of deepfakes on the criminal justice system.

Results and discussion

Figure 1 (see below) shows the number of articles considered at each stage of the systematic review. At the first stage, **1355** articles have been identified through the literature search. After removing **155 duplicates, 1200 were screened for eligibility.** Following this screening stage, which assessed the inclusion criteria and whether it was relevant for the research question, 110 studies met the inclusion criteria for full text-review. Subsequently, 44 studies were included for the analysis.[2]

The findings first examine the geographical focus of research, with most studies on the US. justice system and fewer on the UK and law enforcement. The review then categorises deepfake-related crimes into three main types: pornography and abuse, fraud, and information manipulation and forgery, detailing how each exploits vulnerabilities in digital evidence and courts.

Next, the section highlights the challenges deepfakes create for courts, such as eroding trust in evidence, complicating attribution, and falsifying evidence. The impact on law enforcement is also addressed, noting how deepfakes expand the scope of crime and require advancements in digital forensics and collaboration.

Solutions are outlined, including legislative reforms, technological innovations in detection, and training for legal professionals to better handle deepfake-related issues. The emergence of new cybercrimes facilitated by deepfakes, such as identity theft and misinformation, is also discussed (Figs. 2 and 3).

Threats to the criminal justice system (in itself)

The production, dissemination, and existence of deepfakes present, in itself, a challenge to criminal justice systems across several jurisdictions. Jones and Jones (2022) analysed industry knowledge of deepfakes in the British criminal justice system, and echoing findings by Brookman and Jones (2022), found that the system is inherently vulnerable and unequipped to deal with the threat and its subsequent challenges. This is punctuated by a lack of awareness across the system. The risk is further

[2] See appendix B for final 44 selected studies.

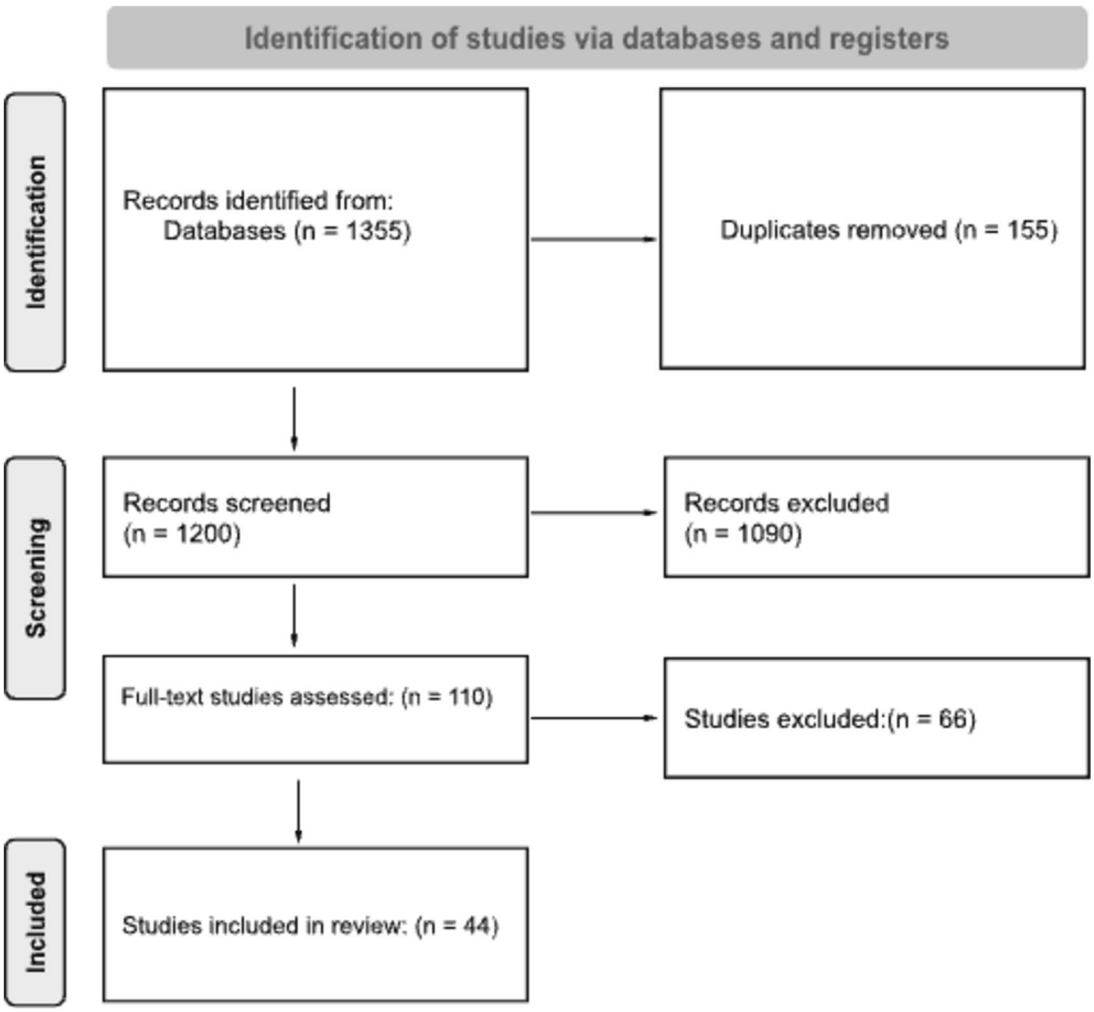

Fig. 1 Search stages and results of the systematic review

exacerbated by deepfake's high accessibility, fast dissemination, and low technical threshold for use (Langa, 2021; Mullen, 2022; Myhand, 2022) This, argues (Ruff, 2021), makes it even harder to fathom the possible impacts deepfakes might have on society and criminal justice systems around the world.

Currently, detection efforts are lagging behind deepfake development and dissemination, making it increasingly difficult to detect them (Myhand, 2022). In courts, however, there are currently no standards, procedures, or rules for addressing this concern. This results in challenges for judges, and lawyers, among others, to ascertain evidence credibility, and upend the respective court processes for evidence verification (Brookman & Jones, 2022).

In the articles analysed in this review, we have identified three main categories of deepfake-related challenges faced by courts:

3.1.1. Erosion of trust and confidence in institutions (legal and law enforcement) and the concept of truth
3.1.2. Issues with attribution and building a legal case
3.1.3. Falsification and modification of evidence in court

Erosion of trust and confidence in institutions (legal and law enforcement) and the concept of truth

Regarding the first point, these types of challenges are a nod to the concept of the theorised 'post-truth' world (Freeman, 2021). In the case of deepfakes, these force the viewer to reevaluate how much trust they put into the videos and images they interact with, especially in court, as they can appear to be true but be false. Even if the viewer knows what they see is false, they can still model perception at a subconscious level. Furthermore,

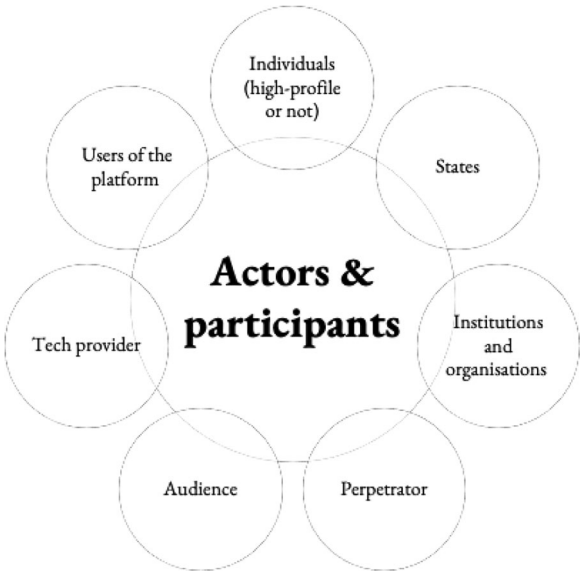

Fig. 2 Eight actors and participants involved in the deepfake operation (Own authorship based on European Parliament. DirectorateGeneral for Parliamentary Research Services. Tackling deepfakes in European policy, 2021)

the mere existence of deepfakes, (Freeman, 2021) finds, erodes this 'inherent' trust placed on video and image-based evidence. Overall, and as a result, deepfakes can lead or contribute to the severe corrosion of justice systems and the rule of law (European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021).

Issues with attribution and building a legal case

The second point shifts the focus to issues faced within the courts. Freeman (Freeman, 2021) highlights the difficulties faced by legal professionals in building cases and establishing attribution in cases of 'traditional' misinformation, especially if deepfake-powered. Furthermore, the legislative gap means professionals struggle to know under which claims to prosecute offences using deepfakes. In the EPRS 2021 report (Ruff, 2021), authors argue deepfakes could, in some cases, be prosecuted under copyright law, highlighting how current civil and criminal laws can be adapted to prosecute deepfakes. Nevertheless, there is no agreement on this point in the literature, with some authors proposing new legislation and more strict regulation of this technology (Cover, 2022; Langa, 2021; Sloot & Wagensveld, 2022a).

Falsification and modification of evidence in court

The third point, when compared to the other categories of challenges, is the most widely discussed in the literature, being one of the most significant threats (Al-Mulla, 2022; Delfino, 2022; Fallis, 2021; Mullen, 2021; Ruff, 2022). As mentioned above, significant levels of trust are put into video and images, and this is no exception in court settings. For instance, when a witness is absent, judges can rely on videos or images as sole 'witnesses'—known as the silent witness theory (Breen, 2021). Usually, standards of verification in these silent witness cases are variable, and in a lot of them, only the evidence's chain of custody is evaluated, but dependent on jurisdictional discrepancies and particularities. This 'inherent' trust is further shown by the fact that the process of authentication

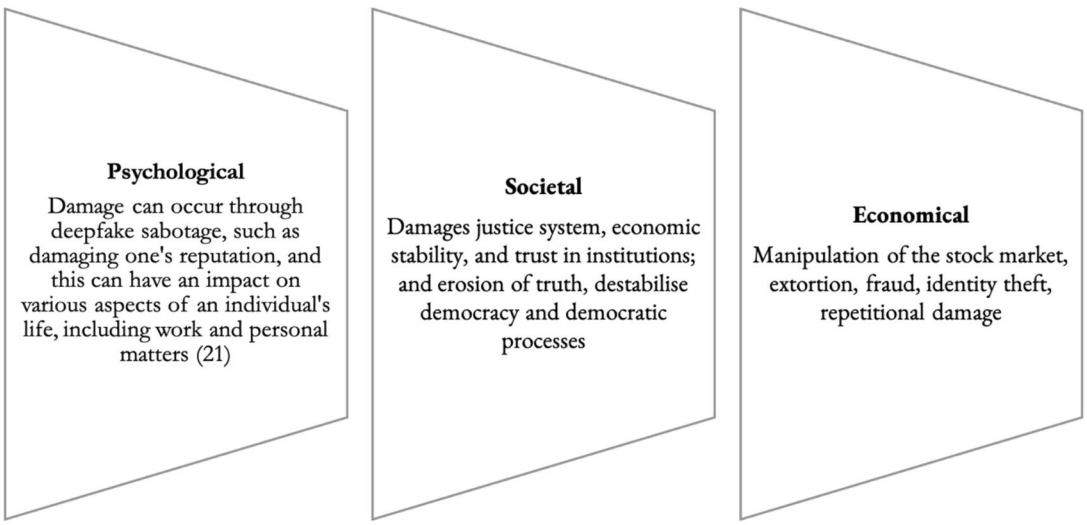

Fig. 3 The impacts of deepfake crimes reflect three primary levels of its effects (Own authorship based on European Parliament.Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021)

happens outside, and thus removed, from the courtroom (Cover, 2022).

Furthermore, deepfakes can overarchingly affect the "(…) credibility and admissibility of audio-visual footage as electronic evidence before the courts" [22, p.58] possibly influencing jurors and testimonies independently of whether they were detected, or not, in a particular instance (Breen, 2021; Cover, 2022; Ruff, 2022). In addition, they are contextualised in a growing online culture, in which fake information or videos are elevated to the level of truth, or near absolute truth (Cover, 2022). Both Ullrich (Ullrich, 2021) and the authors of the EPRS report (European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021) discussed the potential impact of deepfakes on wrongful conditions resulting from fabricated or modified evidence. It's important to note that their discussions are largely based on theoretical considerations and analyses rather than specific evaluations of real cases. These technologies can also influence standards for evidence in cases that are unrelated to deepfakes (Ullrich, 2021). Overall, as a general trend, (Fallis, 2021) argues that video and image-based evidence might start carrying less information and value than what they previously did.

Lastly, the interplay between AI and cognitive biases are usually separately treated in the literature, thus overlooking how AI might reshape the role played by emotions, impressions, feelings, memories, and experience in the judicial approach to complex digital evidence and, consequently, the delivery of criminal justice, further discusses deepfake's impact on suggestibility, as well as their impact on courtroom integrity (Delfino, 2022).

Police

As with courts, law enforcement faces an expanded version of traditional criminal offences and an increased attack surface. The police must, therefore, develop in tandem with the threats, to manage a rapidly evolving and expanding attack surface (Brookman & Jones, 2022; Ligeti, 2019; Yu & Carroll, 2022). Yu and Carroll (2022) and Ligeti (2019) argue that policing these threats must take into consideration four paradoxical pairs:

- Public security and individual rights,
- Transparency of evidence and the black box effect,
- Law enforcement efficiency and legality and
- Emerging crimes and lagging laws

Considering these paradoxes, the process of verification and authentication of criminal evidence and detection by police requires revaluation (Ruff, 2022). This can be done by fostering innovation in digital forensics, expanding the field and implementing more rigorous standards of verification of video, audio and images (Brookman & Jones, 2022). As such, police are faced with three main challenges:

1. Using AI and big data to modernise and build a more robust police force and to assist in detecting and preventing crime in a sustainable, feasible, and ethical manner (Yu & Carroll, 2022).
2. Enhancing national and international collaboration, both intra and inter-industry (Yu & Carroll, 2022).
3. Integrate risk management systems and appropriate data governance and management practices in detection protocols and methods, to prevent privacy violations

Solutions to threats at the court level

Within the literature, several authors highlight early law response and legislation as the key feature in preventing deepfake threats (Cover, 2022; European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021). This has been deemed by Cover (2022) as an alarmist approach, with authors recommending early response on the basis that it will minimise nefarious consequences. Authors falling under and out of the alarmist school of thought all agree that comprehensive solutions are necessary, as this technology is very accessible and spreads quickly and with no regard for jurisdictional boundaries (Langa, 2021).

USA example measures: federal level solutions

Der Sloot and Wagensveld (2022b) propose implementing ex-ante enforcement, to reduce deepfake threats, through means such as market regulation. This could include prohibiting the production, offering, use, or possession of deepfake technology in the consumer market; or introducing an ex-ante legitimacy test, to be carried out before any material is published and/or distributed by citizens or even used in court. However, these measures would result effectively in a ban, and the authors rightfully question how desirable such a ban would be. A ban of this kind raises significant questions surrounding not only its efficacy, but also its ethical, or unethical, nature. Der Sloot and Wagensveld also point out that a ban of this type would most likely not solve the core issues at hand, but rather attempt to contain them.

Alternatively, the authors propose adding additional rules or regulations for statements on public entities, be it persons or groups, as well as for spreading false or misleading information—to tamper with elections or damage

democracy, for example. Regulations of this type are not only productive but also useful and feasible. However, there is a fine line between protecting a public person or limiting the spread of false information and limiting freedom of speech and expression.

Deterring and mitigating

At the level of preventing and mitigating, (Cover, 2022) argues for regulating the production and use of deepfakes in public speaking as well as implementing human moderation as a regulatory deepfakes for deepfakes. On the idea of moderation, Dami (2022) can use blockchain technology to validate content, like videos and images. By using a decentralised, immutable ledger to record information in a way that's continuously verified and re-verified by every entity that uses it, making it nearly impossible to change information after it's been created.

Other authors (Jones & Jones, 2022; Myhand, 2022) focus on education, more specifically of legal professionals, to strengthen the whole justice system and make it more equipped to deal with threats from deepfakes. In the case of the UK, the authors urge the Ministry of Justice to work towards educating legal practitioners about deepfakes, as well as introduce procedures for verifying the authenticity and veracity of evidence used in courts (Cover, 2022).

With regards to jurors in court, Breen (2021) argues the focus of all efforts should be to mitigate doubts surrounding video and image evidence. Some measures could include:

- Using a witness to testify for the accuracy of these videos—based on the pictorial evidence theory, by which a witness testifies the video or image is an accurate representation of what happened or a graphic portrayal of an oral testimony.
- Making changes to silent witness scenarios or cases that rely disproportionately on video or image evidence.
- Establish protocols for checking evidence veracity and authenticity, as well as standards for admitting digital evidence in court (Cover, 2022).
- Adopt tiered categories of probative weights approach, to assess the evidence. This approach is more conscious of information asymmetries, as well as of different jurisdictions. It works by providing a probative assessment of value for different pieces of evidence. Evidence could be confirmed and recognized as true by those submitting it, and if several other sources corroborated this same evidence, then it would hold a high probative value. As categories decreased, the probative value would also decrease, until the last category which would account for non-credible evidence and it would use triangulation for verification and authentication when establishing probative weights (D'Alessandra & Sutherland, 2022).
- Review the process of allocating fact-finding responsibilities for evidence presented in court, as it currently exacerbates the problem and review the chain of decision-making (Freeman, 2021).

Technological responses and detection research

To effectively address the significant threats posed by deepfakes to the criminal justice system, it is crucial to leverage advanced detection technologies. These technological responses not only help in identifying and mitigating deepfakes but also support the judicial process by ensuring the integrity of evidence. By implementing advanced detection methods, the criminal justice system can better verify the authenticity of digital evidence, thus maintaining trust in judicial proceedings. Universities and companies worldwide are developing and researching detection methods to mitigate these threats. Highlighting the contributions of leading institutions and the most effective techniques provides insight into how these advancements support the criminal justice system.

- Leading institutions in deepfake detection research: Universities such as the Chinese Academy of Sciences, Sun Yat-sen University, and Nanjing University are at the forefront of deepfake detection research. These institutions have developed sophisticated algorithms and technologies to identify and mitigate the impact of deepfakes.
- Techniques and technologies: The most popular techniques used in deepfake detection include convolutional neural networks (CNN), machine learning-based techniques, Xception, 3D CNN, and EfficientNet. These technologies focus on detecting inconsistencies in video, audio, and image data that may indicate manipulation.
- Impact of detection research: The advancements in detection technologies are crucial for law enforcement and the judicial system to maintain the integrity of evidence. By identifying deepfakes more effectively, these technologies help ensure that false evidence does not undermine the legal process.

New crimes and threats posed by these crimes

Increasing criminal opportunities exacerbates existing threats or acts as a force multiplier for 'old' crimes. In addition to well-known criminal activities, such as fraud, identity theft, pornography distribution, social

engineering, and automated disinformation, Ali et al. (2022) suggest that this can lead to new kinds of cybercrime and a rise in organised crime. Based on this, the literature shows the significant categorisation of actors involved, types of impacts and the most common variety of deepfakes. The following figures exhibit these categorisations:

Firstly, to understand the deepfake crime operation, eight main actors and participants are identified (European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021):

Secondly, Al-Mulla's (2022) abuse models for classifying the impacts of deepfake crimes reflect three primary levels of its effects (European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021).

As result of, finally the literature stands out three main types of deepfakes crimes discussed:

- Pornography and Abuse
- Fraud
- Information Manipulation and Forgery: Misinformation and Disinformation

Typology and threats of deepfakes crimes

In our review of the literature, we found that authors often considered the expanded attack surface as a threat in itself. Most of the crimes fall under one of the following three categories: pornography and abuse, information manipulation, and fraud. Below is a summary of our findings on how these crimes might pose a threat to the criminal justice system.

Pornography and abuse

Most deepfake technology is used to generate non-consensual pornography, according to Sloot and Wagensveld (European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy, 2021). In fact, a 2019 Deeptrace report (Ligeti, 2019) revealed that 96% of the 14,600 online deepfake videos were used to forge non-consensual pornographic material, with a total of 134,364,438 views across the top four deepfake pornography websites (Ali et al., 2022; Ligeti, 2019). Shockingly, over 90% of deepfakes online are pornographic depictions of women (Ali et al., 2022).

Apart from being used for non-consensual pornography, deepfake technology poses a significant risk for victims of domestic violence as it can be used to threaten and abuse them. (Sloot & Wagensveld, 2022b) EPRS (Ruff, 2022), Hayward & Maas (Ali et al., 2022), Ferreira et al. (Ajder et al., 2019) and Lucas (Deepfake & Fight, 2022) report that the most common uses of deepfakes for abuse include producing non-consensual content, such as revenge, child and fantasy pornography, sextortion (cyber blackmail), and perpetrated abuse. Additionally, Baker et al. (Sloot & Wagensveld, 2022b) suggest that deepfakes are likely to increase the production of these types of illegal pornography.

Moreover, research on the use of deepfakes as a means of technology-facilitated image-based sexual abuse is still in its infancy (Deepfake & Fight, 2022). The potential harm that deepfake technology can cause to individuals and society cannot be overstated, and it is crucial that measures are taken to address the issue.

Deepfakes pose a significant threat due to their automated nature, which supports the hypothetical increase and adds to scale and speed, thus multiplying the threat (Ruff, 2022). Unfortunately, this type of crime is routinely under-reported, as victims tend not to report offences (Popescu, 2021; Ray, 2021). The anonymity behind deepfakes significantly hinders perpetrators' prosecution and victims do not currently have the right to retain anonymity under the law, which contributes to under-reporting trends (Ruff, 2021). Moreover, current laws do not cover these instances and result in a gap for victims and prosecutors alike (Popescu, 2021; Ray, 2021).

Recurring beliefs exacerbate this problem, as non-consensual acts in a digital setting are often viewed as less severe than other forms of abuse and traditional stalking (Chesney, 2019).

Although some progress has been made in the UK, the Law Commission's review began in 2019 and is still underway. Citron [44] argues that privacy injunctions are crucial to prosecute these crimes and address intimate privacy violations effectively.

Furthermore, Henry & Witt (Nagumotu, 2022) identify four overarching shortcomings in the policies and practices of a range of digital platforms in governing image-based sexual abuse: inconsistent, reductionist, and ambiguous language; a stark gap between the policy and practice of content regulation, including transparency deficits; imperfect technology for detecting abuse; and the responsibilities of users to report and prevent abuse. These shortcomings reveal the need for more robust policies and practices to combat the growing threat of deepfakes and image-based sexual abuse.

Fraud

Deepfakes not only enable several forms of financial crimes, but also amplify them (Ruff, 2021). For instance, in 2019, the CEO of a British energy company was asked by fraudsters to transfer AC 220,000 to a bank account in

Hungary using audio deepfake technology to imitate the voice of the CEO of its parent company (Popescu, 2021).

Deepfakes can be utilised in several types of fraudulent activity, like romance fraud (Carvajal Rodriguez, 2021) and grandma scams (Ruff, 2021). They often serve as an entry point or launching pad for other types of crime and are frequently associated with identity theft. This type of offence increases challenges in identity verification and management, which in turn affects society and national security more broadly (Yu & Carroll, 2022). For instance, deepfakes could be used to bypass biometrics in cases of identity theft (Ruff, 2021).

The case of fraud associated with identity theft illustrates the cascading effects of deepfake crimes. It highlights the need for more effective identity verification and management strategies, as well as more robust policies and practices to combat the growing threat of deepfakes.

Information manipulation and forgery: misinformation and disinformation

The increasing realism of deepfakes poses a significant threat to both individual and societal interests, as it can be used to distort democratic discourse, manipulate elections, or jeopardise national security, leading to cascading effects (Brookman & Jones, 2022). According to Ahmed's study (Goodfellow et al., 2014), people who are unaware that they saw a deepfake are more likely to believe its intentions, and the cognitive ability of the exposed individual is linked to ascertaining the sharing intention behind deepfakes.

Moreover, deepfakes can trigger armed conflicts that threaten national security on both sides of the Atlantic, as was the case with the purported deepfake of Gabonese President Ali Bongo Ondimba in 2019. The mere knowledge of deepfakes can undermine the credibility of any video, making it difficult to determine its authenticity, thereby leading to irreparable consequences, whether real or fake (Brookman & Jones, 2021).

Malicious actors can exploit deepfakes to undermine democratic discourse, damage trust in government, and exploit social divisions, which can seriously affect election integrity, military operations, and intelligence gathering (Brookman & Jones, 2021). Criddle (Popescu, 2021) highlights the need to account for these dangers in the UK criminal justice and legal systems, following the example of the EU, which has expanded disinformation laws to account for deepfakes, emphasising the importance of past campaigns as cautionary examples of the significant power that deepfakes wield (37).

Conclusion

To sum up, the criminal justice system is facing significant threats due to the increasing use of deepfakes, making it harder for courts to ascertain the credibility of evidence and for law enforcement agencies to verify the authenticity of the evidence. The production and dissemination of deepfakes have resulted in challenges to the legal and law enforcement institutions, such as the erosion of trust in institutions, issues with attribution and building legal cases, and falsification and modification of evidence in court. Additionally, the police must find ways to balance public security and individual rights, transparency of evidence, law enforcement efficiency and legality, and emerging crimes and lagging laws. The rise in deepfakes has led to new kinds of cybercrime, exacerbating existing threats and having severe psychological, societal, and financial/economic impacts. Pornography and abuse are the most common types of deepfakes crimes discussed in the literature. It is crucial to develop more rigorous standards for verifying video, audio, and images and to foster innovation in digital forensics to address these challenges. National and international collaboration and integrating risk management systems and appropriate data governance and management practices into detection protocols are essential to prevent privacy violations.

To effectively combat these threats to the criminal justice system, it is essential to leverage advanced detection technologies. Although some universities and companies worldwide are developing methods to detect and mitigate deepfakes, the majority of this research is concentrated in China. With 51 universities in China actively engaged in deepfake detection research, compared to 38 universities in the rest of the world combined, this concentration is concerning. It indicates that not all regions are taking the threat seriously enough to invest in necessary research and development. Additionally, the methods used for detection are not diverse in their technology, relying heavily on techniques such as convolutional neural networks (CNN). Given the rapid evolution of deepfake technology, this lack of diversity may minimise the efficiency and effectiveness of detection efforts over time. This lack of widespread, global action exacerbates the risk to the criminal justice system, making it harder to maintain the integrity of judicial processes and evidence.

Legislative Changes: It is crucial to introduce new laws that explicitly address the creation and distribution of deepfakes. Additionally, existing digital privacy laws should be amended to incorporate the challenges posed by deepfakes, ensuring that the legal framework remains robust and relevant in the face of evolving technological threats.

Standards for Evidence Verification: Developing and implementing rigorous standards for verifying digital evidence in court is essential. This includes establishing protocols for verifying the authenticity of videos, audios, and images used in legal proceedings. Such standards will help maintain the integrity of evidence presented in court, ensuring fair and accurate judicial outcomes.

International Collaboration: To effectively combat the global nature of the deepfake threat, fostering national and international collaboration is necessary. Developing and sharing best practices for deepfake detection should be a priority, along with enhancing cooperation between law enforcement agencies, academic institutions, and private sector companies. This collaborative approach will ensure a unified and effective response to deepfake-related challenges.

Education and Training: Providing ongoing training for law enforcement personnel on the latest deepfake detection techniques and the legal implications of using AI-generated evidence is vital. Additionally, educating the public about the risks associated with deepfakes and the importance of digital literacy will help build a more informed and resilient society. Integrating digital literacy programs into educational curricula to raise awareness from an early age will further strengthen this effort.

To stay ahead of the rapidly evolving deepfake technology, it is essential to invest in continuous research and development. Future research should focus on diversifying detection techniques beyond convolutional neural networks, exploring new methodologies and technologies that can improve the accuracy and reliability of deepfake detection. Moreover, interdisciplinary research that combines insights from computer science, law, and social sciences will be crucial in developing comprehensive strategies to combat deepfake threats.

By implementing these recommendations, the criminal justice system can better address the challenges posed by deepfakes and ensure the integrity of its processes and evidence.

Appendix A: Systematic review protocol

I. Title:	The Threat of Deepfakes to the Criminal Justice System: a Systematic Review
II.Summary	Deepfakes are a relatively new phenomenon, surfacing under this same name in 2017. They are algorithmically produced synthetic or altered media, in the format audio, image, video, textual or real-time/live. In more recent years, the term has begun to fall under the broader category of AI-generated synthetic media. An example is face swapping—replacing a person's face with another, in a video or image Deepfakes use deep learning, a form of machine learning, to train generative neural network architectures—the 'tools' used to produce deepfakes. The approaches of deep learning which are more often used to produce deepfakes are GANs (generative adversarial networks) and Autoencoders. Generative Adversarial Networks were first introduced in 2014 by Goodfellow et al. (2014) Deepfakes have gained notoriety in the last 5–6 years. Although some examples of deepfake use are harmless, others, such as a deepfake of President Volodymir Zelensky encouraging Ukrainians surrender at the beginning of the Russian invasion, demonstrate their potential malicious use (Deepfake presidents used in Russia-Ukraine war - BBC News [Internet]., 2022) From a computer science perspective, a significant amount of research has been completed on their creation (Nguyen, 2022) and detection (Pianese et al., 2022). However, significantly less literature is available on their threat to the criminal justice system, which deals with apprehending, prosecuting, defending, sentencing, and punishing those who are suspected or convicted of criminal offences. This systematic review will analyse the literature on the intersection between deepfakes and the criminal justice system and their effect on the latter. This review will also look at what universities and companies are pioneering deepfake detection, as well as what expertise and technologies they use for detection purposes
III. Research questions	RQ: What threats do deepfakes pose to the criminal justice system?

IV. Academic Databases	The following digital databases will be used to identify the literature to be reviewed: 1. Web of Science (Databases covered: Conference Proceedings Citation Index, Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, and Book Citation Index) 2. ProQuest (Databases covered: Library & Information Science Abstracts (LISA), Proquest Central (Criminal Justice Database, Computing Database, Library Science Database, Science Database, Social Science Database, Psychology Database and continent- specific databases covering technology and social sciences (such as Australia & New Zealand Database, Continental Europe Database, East & South Asia Database, East Europe & Central Europe database etc.), ProQuest Dissertations & Theses Global) 3. Scopus (Elsevier's abstract and citation database—Content on Scopus comes from over 5000 publishers and must be reviewed and selected by an independent Content Selection and Advisory Board (CSAB) to be, and continue to be, indexed on Scopus) Another source to be searched is: 4. Google Scholar (Google Scholar includes journal and conference papers, theses and dissertations, academic books, pre-prints, abstracts, technical reports and other scholarly literature from all broad areas of research. It contains works from a wide variety of academic publishers, professional societies and university repositories, as well as scholarly articles available anywhere across the web. Google Scholar also includes court opinions and patents) (Help & [Internet]., 2022)
V. Inclusion criteria for academic literature	Articles identified using a keyword search of the above databases will be assessed using the following criteria: Studies must discuss the threat of deepfakes to the criminal system, particularly vis-a-vis the reliability of evidence in courts. Articles must also focus on threats relevant to the investigation and prosecution stages, including crime and court proceedings, as well as the commissioning of crime Types of papers: - Peer reviewed papers, journal articles, and academic papers - Policy/white papers - Industry reports - Systematic reviews - Meta-analyses - Theses Types of study design to be included: - RCTs - Qualitative and quantitative - Systematic review and meta-analyses - Case reports - Legal documents Articles will be excluded if: - They are published in magazines or newspapers - They are behind a paywall to which UCL does not have access - Master's dissertation For exclusion criteria, we will consider those proposed in (5) and (6), which include: ● Issues with methodological quality ● Review articles with no original data ● Works which are not relevant to the research question and outcomes ● Sources outside the defined time scop (2021–2022) ● Sources in languages other than English
VI. Academic Literature	This systematic review will limit the scope of the search to material published in the UK in 2021–2022 to provide the stakeholder with a recent systematic review. The machine learning model is the common underlying mechanism for deepfakes, although not used in the production of all deepfakes. The term deepfake, however, would only be widely used from 2016–2017 onwards

VII. Search Strategy	The following search terms and search strategies will be used to retrieve the relevant sources for this systematic literature review: "Deepfake*" OR "Deep fake*" OR "Deep-fake*" AND "Court*" OR "Justice system" OR "Proof*" OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim*" OR "trial" Variations of this syntax will be required for different search engines: WEB OF SCIENCE ((TI = ("Deepfake*" OR "Deep fake*" OR "Deep-fake*")) AND TI = ("Court" OR "Justice system" OR "Proof*" OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim*" OR "trial") Scopus "Deepfake*" OR "Deep fake*" OR "Deep?Fake" AND "Justice system" OR "Court" OR "Proof?" OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim*" OR "trial"" Google scholar "Deepfake*" OR "Deep fake* OR Deep-fake*"" AND "Justice system" OR "Court" OR "Proof* OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim* OR "trial"" ProQuest (Deepfake AND ("Justice system" OR "Court" OR "Proof" OR "Evidence" OR "Corroboration" OR "offen*" OR "Threat" OR "crim*" OR "trial"))
VIII. Piloting	Some preliminary piloting has been carried out to assess the adequacy of the search terms. As a part of the piloting, we have adjusted the search terms to encompass relevant and specific articles, whilst also retrieving sources which are more general and loosely related. Key terms have also been adapted to match the databases we will be utilising, including law databases. We consulted with an academic librarian to ensure the appropriate search terms and databases
IX. Literature management	For the academic review, we will use Zotero to manage the literature databases. Zotero is free and open-source software used to manage bibliographic data and related research materials A summary of the volume of articles identified (and excluded) at each stage of the search process will be summarised in a PRISMA flow diagram (39). The diagram depicts the flow of information through the three main stages of the systematic review (see below). It maps out the number of records identified, included and excluded, and the reasons for exclusions Three stages: 1. Identification stage: For stage one, after the removal of duplicates the number of identified articles is established 2. Screening stage: For each paper, the title and abstract will be read to determine if it appears to meet the inclusion criteria. Those that clearly do not, will be excluded at this stage 3. Eligibility stage: All articles that pass stage 2 will be read in full to assess whether they do in fact meet the inclusion criteria
X. Selection of studies	Two researchers will separately read the titles and abstracts of 25% of the identified papers' to assess whether they meet our inclusion criteria (see above). Any inconsistencies will be resolved through discussion and by consulting the other two researchers, if necessary. Where two researchers disagree, a third researcher will be involved. To efficiently screen articles Rayyan will be used—this screening tool is suitable for undertaking systematic review as it makes ordering articles and extracting data convenient Inter-rater reliability (IRR) will be assessed based on two coding categories (i.e. inclusion versus exclusion) using the prevalence- and bias-adjusted kappa (PABAK) statistic, which controls for chance agreement. The following cut-offs will be used to assess IRR: 0.40–0.59 indicates fair agreement, 0.60–0.74 indicates good agreement and > 0.75 indicates high agreement Once adequate inter-rater reliability has been achieved, the remaining 75% of articles will be equally split between two teams of two researchers to be screened. In case conflict arises, one of the members of the other team will review and resolve it All papers that meet the inclusion criteria will be read in full

XI. Data extraction	A pro-forma has been developed by the research team to extract information relevant to the research question. This will be piloted by two researchers on a sample of articles to ensure that relevant information can be captured reliably, and the pro-forma updated as necessary (e.g. we anticipate the need to adjust the terms used to accommodate those employed by research teams from different disciplines). A third researcher will independently check the pro-forma for accuracy and completeness To pinpoint which universities and companies universities/companies have expertise in identifying deepfakes, the names of authors and their respective university affiliations will be extracted. Then, analysis will provide charts with a list of the top 20 researchers and universities with a count of the number of publications, as well as their specific types of expertise namely audio, image, video deepfake detection Pro-forma details to be collated: • Year of study • Publication type (journal, paper in conference proceedings etc.) • University affiliation • Author • Consequences/impact on and to: - Legal proceedings with new types of persecution and/or trials - Authenticity of evidence in Courts - New tools to commit crime such as fraud or child pornography - Risk of wrongly convicting the defendant of a crime
XII.Method of synthesis	As a method of synthesis, thematic analysis will be used to systematically identify, organise, and provide insight into patterns of meaning across the data set (Braun and Clarke, 2012)
XIII.Target journal	Journal of Cybersecurity/Crime Science Journal

Appendix B: 44 papers selected for the final analysis

1. Ali, A., Khan Ghouri, K. F., Naseem, H., Soomro, T. R., Mansoor, W., & Momani, A. M. (2022). Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security. The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings. https://doi.org/https://doi.org/10.1109/ICCR56254.2022.9995821

2. Allen, D. (2021). Deepfake Fight: AI-Powered Disinformation and Perfidy Under the Geneva Conventions (SSRN Scholarly Paper 3958426). https://doi.org/https://doi.org/10.2139/ssrn.3958426

3. Al-Mulla, M. S. (2022). Deepfakes: Criminalization And Legalization Analytical Descriptive Study. Webology, 19(2), 3210–3223. https://www.proquest.com/docview/2695094996/abstract/EE666980E3414A46PQ/1

4. Atif, Babiker, Mohamed, Ali, Khushboo, Farid, Khan, Ghouri., Hina, Naseem, Tariq, Rahim, Soomro, Wathiq, Mansoor., & Al, Momani. (2022). Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security. SciSpace—Paper, 1–5. https://doi.org/https://doi.org/10.1109/ICCR56254.2022.9995821

5. Brookman, F., & Jones, H. (2022). Capturing killers: The construction of CCTV evidence during homicide investigations. Policing and Society, 32(2), 125–144. https://doi.org/https://doi.org/10.1080/10439463.2021.1879075

6. Carvajal Rodriguez, L. C. (2021). Political Deepfakes: Cultural Discourses of Synthetic Audio-Visual Manipulations [M.A.]. https://www.proquest.com/docview/2541745984/abstract/57DEF4AFCAFF4737PQ/1

7. Citron, D. (2022). Privacy Injunctions. Emory Law Journal, 71(5), 955. https://scholarlycommons.law.emory.edu/elj/vol71/iss5/3

8. Cover, R. (2022). Deepfake culture: The emergence of audio–video deception as an object of social anxiety and regulation. Continuum, 36(4), 609–621. https://doi.org/https://doi.org/10.1080/10304312.2022.2084039

9. Criddle, C. (2022). Call for sharing of deepfake porn to be made illegal in the UK. FT.Com. https://www.proquest.com/docview/2698559311/citation/BCFCE268CDBA4B54PQ/1

10. Criddle, C., Cameron-Chileshe, J., & Johnston, I. (2022). UK government drops 'legal but harmful' clause from new online law. FT.Com. https://www.proquest.com/docview/2758525125/citation/D84B003B149846FAPQ/1

11. Dami, L. (2022). Analysis and conceptualization of deepfake technology as cyber threat. https://doi.org/https://doi.org/10.13140/RG.2.2.21862.50246

12. Danielle C. Breen, J. D. (2021). Silent No More: How Deepfakes Will Force Courts To Reconsider Video Admission Standards. Journal Of High Technology Law, Vol. Xxi: No. 1. https://bpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/5/1153/files/2021/01/Breen.pdf

13. de Rancourt-Raymond, A., & Smaili, N. (2022). The unethical use of deepfakes. Journal of Financial Crime, ahead-of-print (ahead-of-print). https://doi.org/https://doi.org/10.1108/JFC-04-2022-0090

14. European Parliament. Directorate General for Parliamentary Research Services. (2021). Tackling deepfakes in European policy. Publications Office. https://data.europa.eu/doi/https://doi.org/10.2861/325063

15. Fallis, D. (2021). The Epistemic Threat of Deepfakes. Philosophy & Technology, 34(4), 623–643. https://doi.org/https://doi.org/10.1007/s13347-020-00419-2

16. Ferreira, S., Antunes, M., & Correia, M. E. (2021). Exposing Manipulated Photos and Videos in Digital Forensics Analysis. Journal of Imaging, 7(7), Article 7. https://doi.org/https://doi.org/10.3390/jimaging7070102

17. Freeman, L. (2021a). Hacked and Leaked: Legal Issues Arising From the Use of Unlawfully Obtained Digital Evidence in International Criminal Cases. UCLA Journal of International Law and Foreign Affairs, 25(2). https://escholarship.org/uc/item/5b87861x

18. Freeman, L. (2021b). Weapons of War, Tools of Justice: Using Artificial Intelligence to Investigate International Crimes. Journal of International Criminal Justice, 19(1), 35–53. https://doi.org/https://doi.org/10.1093/jicj/mqab013

19. Harris, K. R. (2021). Video on Demand: What Deepfakes Do and How They Harm. Synthese, 199(5–6), 13373–13391. https://doi.org/https://doi.org/10.1007/s11229-021-03379-y

20. Hayward, K. J., & Maas, M. M. (2021). Artificial intelligence and crime: A primer for criminologists. Crime, Media, Culture, 17(2), 209–233. https://doi.org/https://doi.org/10.1177/1741659020917434

21. Huijstee, M. van, Boheemen, P. van, Das, D., Nierling, L., Jahnel, J., Karaboga, M., & Fatun, M. (2021). Tackling Deepfakes in European Policy. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

22. Jamie Baker, Laurie Hobart, & Matthew Mittelsteadt. (2021). AI for Judges. https://cset.georgetown.edu/wp-content/uploads/AI-for-Judges.pdf

23. Jones, K. O., & Jones, B. S. (2022). How robust is the United Kingdom justice system against the advance of deepfake audio and video? https://epluse.ceec.bg/how-robust-is-the-united-kingdom-justice-system-against-the-advance-of-deepfake-audio-and-video/

24. Kavyasri Nagumotu. (2022). Deepfakes are Taking Over Social Media: Can the Law Keep Up? IDEA – The Journal of the Franklin Pierce Center for Intellectual Property, Volume 62 – Number 2. https://cdn.ymaws.com/www.bpla.org/resource/resmgr/docs/Nagumotu_Paper.pdf

25. Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation To Combat Threats Posed By Deepfakes. Boston University Law Review, 101. https://www.bu.edu/bulawreview/files/2021/04/LANGA.pdf

26. Lewis, C. (2022). The Need for a Legal Framework to Regulate the Use of Artificial Intelligence

27. Ligeti, K., & Robinson, G. (2021). The Handling of Digital Evidence in Luxembourg. https://orbilu.uni.lu/handle/10993/45164

28. Lucas, K. T. (2022). Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology. Victims & Offenders, 17(5), 647–659. https://doi.org/https://doi.org/10.1080/15564886.2022.2036656

29. Mullen, Molly. (2022). A New Reality: Deepfake Technology and the World Around Us. New Scientist, 192(2578), 61. https://doi.org/https://doi.org/10.1016/S0262-4079(06)61134-2

30. Myhand, T. (2022). Once the Jury Sees it, the Jury Can't Unsee it: The Challenge Trial Judges Face When Authenticating Video Evidence in the Age of Deepfakes. SSRN Electronic Journal. https://doi.org/https://doi.org/10.2139/ssrn.4270735

31. Nagumotu, K. (2022). DEEPFAKES ARE TAKING OVER SOCIAL MEDIA: CAN THE LAW KEEP UP? 62(2)

32. Popescu, M. M. (2021). Disinformation Deconstructedcognition Security and Digital Control. Bulletin of the Transilvania University of Brasov. Series VII, Social Sciences and Law., 14(1), 123–134. https://doi.org/https://doi.org/10.31926/but.ssl.2021.14.63.1.12

33. Ray, A. (2021). Disinformation, Deepfakes and Democracies: The Need for Legislative Reform. University of New South Wales Law Journal, 44(3). https://doi.org/https://doi.org/10.53637/DELS2700

34. Ristovska, S. (2022). Deepfakes and Their (Un)intended Consequences. Scitech Lawyer, 19(1), 12–16. https://www.proquest.com/docview/2738616937/abstract/3EB33523046B4897PQ/1

35. Ruff, J. C. (2021). The Federal Rules of Evidence Are Prepared for Deepfakes. Are You? The Review of Litigation, 41(1), 103–126. https://www.proquest.com/docview/2635270713/abstract/304789636F654A12PQ/1

36. Segate, R. V. (2021). Cognitive Bias, Privacy Rights, and Digital Evidence in International Criminal Proceedings: Demystifying the Double-Edged ai Revolution. International Criminal Law Review, 21(2), 242–279. https://doi.org/https://doi.org/10.1163/15718123-bja10048

37. Silva, R. B. da. (2021). Updating the Authentication of Digital Evidence in the International Criminal Court. International Criminal Law Review, 22(5–6), 941–964. https://doi.org/https://doi.org/10.1163/15718123-bja10083

38. Suslavich, B. T. (2022). Nonconsensual Deepfakes: A 'Deep Problem' For Victims. Journal of Science and Technology, 29. https://www.albanylawscitech.org/article/74850-nonconsensual-deepfakes-a-deep-problem-for-victims

39. Todd C. Helmus. (2022). Artificial Intelligence, Deepfakes, and Disinformation: A Primer. RAND Corporation. https://doi.org/https://doi.org/10.7249/PEA1043-1

40. Ullrich, Q. J. (2021). Is This Video Real? The Principal Mischief of Deepfakes and How the Lanham Act Can Address It. Columbia Journal of Law and Social Problems, 55(1), 1–56. https://www.proquest.com/docview/2629434387/abstract/9AEE37465A8F4212PQ/1

41. Van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. Computer Law & Security Review, 46, 105716. https://doi.org/https://doi.org/10.1016/j.clsr.2022.105716

42. Vasist, P. N., & Krishnan, S. (2022). Engaging with deepfakes: A meta-synthesis from the perspective of social shaping of technology theory. Internet Research, ahead-of-print (ahead-of-print). https://doi.org/https://doi.org/10.1108/INTR-06-2022-0465

43. Vincenzo Ciancaglini, Craig Gibson, David Sancho, Odhran McCarthy, Maria Eira, Philipp Amann, & Aglika Klayn. (2021). Malicious Uses and Abuses of Artificial Intelligence. https://unicri.it/sites/default/files/2020-11/Abuse_ai.pdf

44. Yu, S., & Carroll, F. (2022). Insights into the Next Generation of Policing: Understanding the Impact of Technology on the Police Force in the Digital Age. In R. Montasari (Ed.), Artificial Intelligence and National Security (pp. 169–191). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-031-06709-9_9

Acknowledgements
We acknowledge Professor Shane D Johnson for their assistance in forming this systematic review.

Author contributions
All authors read and approved the final manuscript.

Funding
This project was funded by the UK EPSRC grant [EP/S022503/1] that supports the Centre for Doctoral Training in Cybersecuritydelivered by UCL's Departments of Computer Science, Security and Crime Science, and Science, Technology, Engineering, andPublic Policy.

Availability of data and materials
Not applicable

Declarations

Competing interests
The authors declare that they have no competing interests. We have no known conflict of interest to disclose.

Received: 18 January 2024 Accepted: 31 October 2024
Published online: 17 November 2024

References

Ajder, H., Patrini, G., Cavalli, F., Cullen, L. (2019). The State of Deepfakes: Landscape, Threats, and Impact [Internet]. Available from: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

Ali A, Khan Ghouri KF, Naseem H, Soomro TR, Mansoor W, Momani AM. Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security. Inst Electr Electron Eng Inc IEEE Conf Proc [Internet]. 2022; Available from: https://www.proquest.com/conference-papers-proceedings/battle-deep-fakes-artificial-intelligence-set/docview/2760666019/se-2?accountid=14511

Allen, D. (2021). Deepfake Fight: AI-Powered Disinformation and Perfidy Under the Geneva Conventions. Available SSRN 3958426 [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3958426.

Al-Mulla, M.S. (2022). Deepfakes: Criminalization And Legalization Analytical Descriptive Study. Webology [Internet]. 19(2):3210–23. Available from: https://www.proquest.com/scholarly-journals/deepfakes-criminalization-legalization-analytical/docview/2695094996/se-2?accountid=14511.

Bajpai, S., Bajpai, R., & Chaturvedi, H. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology, 12*(41), 20–27.

Breen, D. (2021). Silent no more: How deepfakes will force courts to reconsider video admission standards. Journal of High Technology Law [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jhtl21§ion=6.

Brookman, F., Jones, H. (2022). Capturing killers: the construction of CCTV evidence during homicide investigations. Polic Soc [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://doi.org/10.1080/10439463.2021.1879075.

Carvajal Rodriguez, L.C. (2021). Political Deepfakes: Cultural Discourses of Synthetic Audio-Visual Manipulations [Internet] [M.A.]. ProQuest Dissertations and Theses. [Ann Arbor]: Temple University. Available from: https://www.proquest.com/dissertations-theses/political-deepfakes-cultural-discourses-synthetic/docview/2541745984/se-2?accountid=14511.

Chesney, R. (2019). Deepfakes and the new disinformation war. Vol. 98, Foreign Affairs. p. 147–55.

Cover, R. (2022). Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. Continuum [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://doi.org/10.1080/10304312.2022.2084039.

D'Alessandra F, Sutherland K. The promise and challenges of new actors and new technologies in international justice. International Criminal Justice [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://academic.oup.com/jicj/article-abstract/19/1/9/6294452. Deepfake presidents used in Russia-Ukraine war—BBC News. (n.d.). Retrieved 16 November 2022, from https://www.bbc.co.uk/news/technology-60780142.

Dami, L. (2022). Analysis and conceptualization of deepfake technology as cyber threat.

Deepfake presidents used in Russia-Ukraine war - BBC News [Internet]. [cited 2022 Nov 16]. Available from: https://www.bbc.co.uk/news/technology-60780142

Delfino, R. (2022). Deepfakes on Trial: a Call to Expand the Trial Judge'S Gatekeeping Role to Protect Legal Proceedings from Technological Fakery. Available SSRN 4032094 [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4032094.

Edanz-Learning-Team. Understanding a PRISMA Flow Diagram [Internet]. Edanz Learning Lab. 2022 [cited 2022 Dec 13]. Available from: https://learning.edanz.com/prisma-flow-diagram/

European Parliament. Directorate General for Parliamentary Research Services. Tackling deepfakes in European policy. [Internet]. LU: Publications Office; 2021 [cited 2023 Mar 16]. Available from: https://doi.org/10.2861/325063

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy Technology, 34*(4), 623–643.

Fink, A. (2010). Survey Research Methods. In: Peterson P, Baker E, McGaw B, editors. International Encyclopedia of Education (Third Edition) [Internet]. Oxford: Elsevier; [cited 2023 Mar 22]. p. 152–60. Available from: https://www.sciencedirect.com/science/article/pii/B9780080448947002967.

Freeman, L. (2021). Weapons of War, Tools of Justice: Using Artificial Intelligence to Investigate International Crimes. Journal of International Criminal Justice [Internet]. [cited 2023 Mar 21];19(1):35–53. Available from: https://doi.org/10.1093/jicj/mqab013.

Godfellow, I. et al. (2014). Generative Adversarial Networks. Available from: https://arxiv.org/abs/1406.2661.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S, et al. (2014). Generative Adversarial Networks [Internet]. arXiv [cited 2022 Nov 16]. Available from: http://arxiv.org/abs/1406.2661.

Google Scholar Search Help [Internet]. [cited 2022 Dec 15]. Available from: https://scholar.google.com/intl/us/scholar/help.html#coverage

Jones, K., Jones, B. (2022). How robust is the United Kingdom justice system against the advance of deepfake audio and video? 36th Int Conf [Internet]. (Query date: 2023–01–11 17:07:35). Available from: http://insight.cumbria.ac.uk/id/eprint/6675/.

Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *Boston University Law Review, 101*(2), 761–801.

Ligeti, K. (2019). Artificial Intelligence and Criminal Justice.

Malicious Uses and Abuses of Artificial Intelligence [Internet]. Europol. [cited 2023 Mar 21]. Available from: https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence

Meline, T. (2006). Selecting Studies for Systemic Review: Inclusion and Exclusion Criteria. Contemp Issues Commun Sci Disord [Internet]. Mar [cited 2022 Dec 13];33(Spring):21–7. Available from: https://doi.org/10.1044/cicsd_33_S_21.

Mullen M. A New Reality: Deepfake Technology and the World around Us. Mitchell Hamline Rev [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/wmitch48§ion=7.

Myhand, T. (2022). Once the Jury Sees it, the Jury Can't Unsee it: The Challenge Trial Judges Face When Authenticating Video Evidence in the Age of Deepfakes. Available SSRN 4270735 [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4270735.

Nagumotu, K. (2022). Deepfakes are Taking Over Social Media: Can the Law Keep Up? IDEA– The Law Review of the Franklin Pierce Center for Intellectual Property [Internet]. [cited 2023 Feb 23];Volume 62 – Number 2. Available from: https://cdn.ymaws.com/www.bpla.org/resource/resmgr/docs/Nagumotu_Paper.pdf.

Nguyen, T.T. (2022). Deep learning for deepfakes creation and detection: A survey. Comput Vis Image Underst [Internet]. 223(Query date: 2023–01–11 17:20:02). Available from: https://api.elsevier.com/content/article/eid/1-s2.0-S1077314222001114.

Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ [Internet]. Mar 29 [cited 2022 Dec 13];n160. Available from: https://doi.org/10.1136/bmj.n160.

Pianese, A., Cozzolino, D., Poggi, G., Verdoliva, L. (2022). Deepfake audio detection by speaker verification. arXiv.org [Internet]. Sep 28; Available from: https://www.proquest.com/working-papers/deepfake-audio-detection-speaker-verification/docview/2719225671/se-2.

Popescu, M.M. (2021a). Disinformation Deconstructedcognition Security and Digital Control. Bull Transilv Univ Brasov Ser VII Soc Sci Law [Internet]. [cited 2023 Mar 22];14(1):123–34. Available from: https://www.proquest.com/docview/2569414461/abstract/44BB725B1DEE4013PQ/1.

Popescu, M.M. (2021b). Disinformation Deconstructedcognition Security And Digital Control. Bull Transilv Univ Brasov Ser VII Soc Sci Law [Internet]. 14(1):123–34. Available from: https://www.proquest.com/scholarly-journals/disinformation-deconstructedcognition-security/docview/2569414461/se-2.

Ray, A. (2021). Disinformation, deepfakes and democracies: The need for legislative reform. *The University of New South Wales Law Journal, 44*(3), 983–1013.

Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making - Jyoti Belur, Lisa Tompson, Amy Thornton, Miranda Simon, 2021 [Internet]. [cited 2023 Mar 21]. Available from: https://doi.org/10.1177/0049124118799372

Ruff J. The federal rules of evidence are prepared for deepfakes. Are you? Rev Litig [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://search.proquest.com/openview/84ea49671c7f8712bb4b6cef482a6a51/1?pq-origsite=gscholar&cbl=37465.

van der Sloot, B., Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. The Computer Law and Security [Internet]. [cited

2023 Feb 23];46:105716. Available from: https://www.sciencedirect.com/science/article/pii/S0267364922000632.
Sloot, B, Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Comput LAW and Security Review.
Ullrich, Q.J. (2021). Is This Video Real? The Principal Mischief of Deepfakes and How the Lanham Act Can Address It. Columbia Journal of Law and Social Problems [Internet]. Fall;55(1):1–56. Available from: https://www.proquest.com/scholarly-journals/is-this-video-real-principal-mischief-deepfakes/docview/2629434387/se-2?accountid=14511.
Yu, S., Carroll, F. (2022). Insights into the Next Generation of Policing: Understanding the Impact of Technology on the Police Force in the Digital Age. Artificial Intelligence Security Center [Internet]. (Query date: 2023–01–11 17:07:35). Available from: https://doi.org/10.1007/978-3-031-06709-9_9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.