Maras and Logie *Crime Science* (2024) 13:31
https://doi.org/10.1186/s40163-024-00226-6

Crime Science

THEORETICAL ARTICLE Open Access

Countering the complex, multifaceted nature of nude and sexually explicit deepfakes: an Augean task?

Marie-Helen Maras[1*] and Kenji Logie[2]

Abstract

Manipulated media – images, video, and audio – have proliferated online. Widely available modification software, rudimentary practices, and advanced techniques involving machine learning and artificial intelligence have been used to manipulate media. This article critically explores illegal and nonconsensual nude and sexually explicit (NSE) deepfakes and the measures implemented to counter them. The motivating questions for this analysis are: What measures are in place to counter illegal and nonconsensual NSE deepfakes? Are these measures sufficient? The objectives of this article are three-fold: (1) to identify illegal and nonconsensual NSE deepfakes and illegal uses of them; (2) to critically evaluate the current legal and technological countermeasures available in various jurisdictions to combat illicit and nonconsensual NSE deepfakes; and (3) to make recommendations based on challenges and deficits in existing legal and technological mechanisms employed to tackle them. Ultimately, our findings indicate that there are technologies and legal measures that could effectively reduce the harm experienced by victims if: there is a collective will from the corporate, legislative, and political spheres to effectively execute these changes; and legal liability to remove NSE deepfakes is placed on the online platforms and websites that host and distribute this content.

Keywords Deepfake, Cybercrime, Technology-facilitated gender-based violence, Child sexual abuse material

Introduction

Manipulated media are not a new phenomenon. Manipulated media – images, video, and audio – have been created and distributed online to cause personal, social, economic, and political harm (Appel & Prietzel, 2022; Hancock & Bailenson, 2021). Media is manipulated using widely available modification software, rudimentary practices, and advanced techniques involving machine learning and artificial intelligence. This article focuses on one form of manipulated media, namely deepfakes. Deepfakes are generated media or manipulated media using artificial intelligence or machine learning with varying degrees of sophistication to produce highly realistic media portraying real individuals saying or doing things they did not actually say or do. The material used to create deepfakes can be directly or indirectly obtained from various sources, including from the individuals depicted in the material, others with access to that material, and information and communication technologies from users' passive and active digital footprints. Deepfake applications are trained on a media dataset of the

*Correspondence:
Marie-Helen Maras
mmaras@jjay.cuny.edu
[1]Department of Security, Fire, and Emergency Management, and Center for Cybercrime Studies, John Jay College of Criminal Justice, City University of New York, 524 W. 59th Street, Haaren Hall, Room 43311, New York, NY 10019, USA
[2]John Jay College of Criminal Justice, City University of New York, New York, USA

victim to create computer-generated media content by altering, merging, and/or superimposing existing images, videos, and/or audio clips to create a media file that in some cases can easily be mistaken as authentic (Maras & Alexandrou, 2019).

The legality of deepfakes depends on the jurisdiction, the content of the deepfakes, and their use. Deepfakes have been used against adults and children to perpetrate various cybercrimes relating to intellectual property crime, child sexual abuse material, fraud, and technology-facilitated violence, such as cyberharassment, cyberstalking, sextortion, and image-based sexual abuse, among other illicit acts. Nevertheless, this form of manipulated media predominantly depicts and targets women.

This article critically explores the role of nude and sexually explicit deepfakes (hereafter NSE deepfakes) in various cybercrimes and the measures implemented to counter them. The motivating questions for this analysis are: What measures are in place to address the use of illegal and nonconsensual NSE deepfakes to commit cybercrime? Are these measures sufficient to counter them? The article's objectives are three-fold: (1) to identify illegal and nonconsensual NSE deepfakes; (2) to critically assess laws and technologies in various jurisdictions that can be used to counter them; and (3) to propose measures that can deter and minimize the creation and distribution of illegal and nonconsensual NSE deepfakes.

Deepfakes: in brief

Machine learning (ML) and artificial intelligence (AI) applications and programs have been used to manipulate media, which have been described as 'deepfakes.' The term is a "portmanteau of 'deep learning' and 'fake media,'" and has its origins in Reddit ("r/deepfakes") (Trend Micro, UNICRI, EC3, 2020). Deepfakes can include manipulated audio, images, and/or video. The tools used to create deepfakes may replace or re-enact something depicted in media (e.g., swap a face with another face or manipulate facial features), generate something new (e.g., creating a fictional person), and/or synthesize content (e.g., create audio content) (Trend Micro, UNICRI, EC3, 2020). Some tools require a significant amount of training material (e.g., DeepFaceLab) to produce a good quality deepfake (Suwajanakorn et al., 2017); however, others require less training material (e.g., FaceSwap and AttGAN) to produce a quality deepfake (Averbuch-Elor et al., 2017; Fried et al., 2019; He et al., 2019). Deepfake companies are not always forthcoming about their collection, distribution, and retention of training material. For example, while one deepfake company known as FaceApp obscured its cross-border data transfer, data processing, and data retention practices (Maras & Logie, 2021), another company, DeepFaker, was transparent about its ability to collect users' facial features, images, and videos when the app was in use (DeepFaker App, 2022). The adverse implications of the collection, sharing, and retention of deepfake training material on privacy and data protection has been explored by researchers and data protection organizations (e.g., Maras & Logie, 2021; CEDPO, 2023).

Tools to create deepfakes are widely available for free or for a fee and require users to upload media (i.e., video, audio and/or images) to train the ML/AI algorithm to create deepfakes. An example of this tool is the Zao Chinese smartphone application, which enables users to "become movie stars" by superimposing users' faces on the bodies of actors or actresses in films (Doofman, 2019). While the Zao app demonstrates a benign use of deepfakes, deepfakes have also been used for degrading, disturbing, and illegal purposes.

Deepfakes are not the only type of media manipulation that can cause harm. Shallowfakes (or cheapfakes), audio-visual media that is manipulated using less sophisticated technology, such as basic editing tools, can also cause significant harm. This type of low-tech manipulation occurred with Emma González, a survivor of the Douglas High School shooting. She and other students became part of a gun-control movement, March for Our Lives. A Teen Vogue photo of her ripping a paper shooting target was altered to depict her ripping a copy of the U.S. Constitution to paint her in a false light and discredit her (Mezzofiore, 2018). In 2019 and then again in 2020, the audio of a video of the (then) U.S. House Speaker, Nancy Pelosi, was slowed down making it seem as if she was slurring her speech (Reuters, 2020; Sadiq, 2019). This cheapfake still remains online and many distribute it as a statement of fact, even though it has been debunked by many legitimate sources (e.g., Reuters, 2020; Sadiq, 2019).

What makes deepfakes illegal

The illegality of NSE deepfakes depends on their content and the way they are used. Deepfakes can be used to commit cyber-enabled crime (i.e., crime whereby technology plays an essential role in facilitating the crime), including content-related crime (i.e., illegal content). The next sections examine NSE deepfakes that are a specific form of content-related crime, child sexual abuse material, and NSE deepfakes that are used to commit specific forms of cyber-enabled crime, particularly technology facilitated gender-based violence.

Child sexual abuse material (CSAM)

Deepfakes that depict child sexual abuse material (colloquially known as child pornography[1]) are

[1] The term child pornography is inappropriate as there is no element of consent and the use of the term 'child pornography' does not accurately depict the nature of the crime – the sexual abuse of children.

a content-related crime. Many countries criminalize some facet of technology-facilitated CSAM offenses (for information about countries' laws, see ICMEC, 2023). In the United States, a federal law, the Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today (PROTECT) Act of 2003, criminalizes computer-generated child sexual abuse material (18 U.S.C. § 2256(8)). Nevertheless, there are only a few states that currently either specifically criminalize CSAM deepfakes (e.g., Louisiana; see Louisiana Revised Statutes § 14:73.13.C.1) or the unlawful production or distribution of NSE deepfakes (e.g., Texas; see Texas Penal Code § 21.165). In one of the U.S. states that explicitly prohibits its CSAM deepfakes, Louisiana, a suspect was charged under their 2023 deepfake law for creating CSAM deepfakes (Vincent, 2023).

Many laws require that a real child be depicted in CSAM. For example, in the U.S., certain states require that a real child be depicted in synthetic media (e.g., "identifiable child" requirement in Idaho Code § 18-1507, and "identifiable minor" requirement in enacted Iowa Senate Bill 2243, Washington House Bill No. 1999, and Utah House Bills No. 148 and 238). Laws like these fall short of criminalizing the creation of deepfakes that contain realistic depictions of children that do not depict a single, real child, but are often designed with tools that are trained on materials featuring children that serve as the basis to create synthetic media that depict a *virtual child* (i.e., a virtual representation of a child that does not depict an actual or identifiable minor). Due to gaps in existing legislation in the United States, in 2023, the Attorney Generals in all U.S. states expressed concerns about their ability to deal with AI-generated CSAM and asked the U.S. Congress to expand existing laws to cover AI-generated CSAM (Kinnard, 2023). In 2024, certain U.S. states updated their laws to include the criminalization of virtual representations of non-identifiable children (i.e., virtual children). For instance, in 2024, Oklahoma enacted House Bill No. 3642, which criminalizes "[a]ny visual depiction that appears to be a child, regardless of whether the image is a depiction of an actual child, a computer-generated image, or an image altered to appear to be a child, engaged in any act of sexually explicit conduct…" Beyond U.S. state law, federal obscenity law has been used against CSAM deepfakes. Particularly, in May 2024, an existing federal obscenity law, 18 U.S.C. § 1466 A ("[o]bscene visual representations of the sexual abuse of children"), was used to charge a Wisconsin man for creating, distributing, and possessing with the intent to distribute thousands of CSAM deepfakes (*United States of America v. Steven Anderegg*, 2024b). Specifically, according to a court document, he used "generative artificial intelligence ("GenAI") to create hyperrealistic images of nude and semi-clothed prepubescent children lasciviously exhibiting or touching their genitals or engaging in sexual intercourse" (*United States of America v. Steven Anderegg*, 2024a, p. 1) and posted these images on Instagram and advertised their availability on his Telegram channel (Del Valle, 2024).[2] Other countries consider sexually explicit material depicting virtual representations of children as CSAM. For instance, in the Republic of Korea, a court's ruling extended the "legal definition of sexually exploitative material" to include depictions of virtual children (Bae & Yeung, 2023).

Countries that do not have laws that specifically criminalize CSAM created using deepfake software and technology can still investigate and prosecute cases that involve them, if their CSAM laws are broad enough to include any form of CSAM, regardless of the means used to create and/or manipulate CSAM (by, for example, including language like 'any form,' 'any means,' 'other means,' or 'irrespective of medium' in the laws) (for laws covering 'technology-facilitated CSAM offenses,' see ICMEC, 2023) and, for the depiction of virtual children, do not require the depiction of a real child in CSAM deepfakes to proscribe their development, distribution, and possession.

Technology-facilitated gender-based violence

Information and communication technology (ICT) is used to commit a variety of cybercrimes against particular genders worldwide. Women and girls are the predominant targets of technology-facilitated violence, which "is [considered] any act that is committed, assisted, aggravated or amplified by the use of information communication technologies or other digital tools which results in or is likely to result in physical, sexual, psychological, social, political or economic harm or other infringements of rights and freedoms" (UN Women, 2023). NSE deepfakes are a "gendered form of abuse" as the content is "predominantly produced by and for male audiences" (Öhman, 2020, p. 134) and disproportionately depicts and targets women (Adjer et al., 2019; Wang & Kim, 2022). While there are certain jurisdictions where the development, distribution, and promotion of pornography (i.e., sexually explicit content) is considered illegal (e.g., Jordan's Cybercrime Law No. 17 of 2023; Saudi Arabia, Anti-Cyber Crime Law of 2007; Tanzania, Cyber Crimes Act 2015), many jurisdictions do not consider pornography illegal content and do not ban it outright. NSE deepfakes are used to commit various forms of technology-facilitated violence against women, including (but not limited to) image-based sexual abuse, cyberbullying, cyberharassment, sextortion, invasions of privacy, and defamation (The Economist Intelligence Unit, 2021;

[2] The trial of this person is pending as of June 2024.

van der Wilk, 2021; van der Sloot & Wagensveld, 2022; O'Brien & Maras, 2024).

Image-based sexual abuse

Research has shown that the majority of deepfakes online are NSE deepfakes depicting women. One company, Deeptrace Labs, reported in 2019, that most deepfakes (about 96% of those identified) were pornographic, 46% of which depicted female celebrities in the United States and the United Kingdom, and 25% of them depicted female K-pop celebrities (Adjer et al., 2019; Wang & Kim, 2022). 'Nonconsensual pornography' is a term used to describe the distribution of nude and/or sexually explicit images, videos, and/or other forms of media of a person without their consent. 'Nonconsensual pornography' has been used synonymously with the term 'revenge pornography.' However, neither 'nonconsensual pornography' nor 'revenge pornography' accurately describes the crime that is committed and what the victim experiences. The preferred term is 'image-based sexual abuse' (IBSA) which encompasses the creation, distribution, and the threatening to distribute of nude and/or sexually explicit images, videos, and other forms of media (McGlynn et al., 2017; Henry et al., 2019), including manipulated media, such as transposing the victim's face onto bodies depicted in nude and sexually explicit media to create NSE deepfakes. NSE deepfakes are used to degrade, embarrass, and silence women and cause them significant harm (Turk, 2019). An investigative journalist in India became the target of continued and inescapable harassment and abuse online, even being depicted in NSE pornography, to silence her after speaking up following the rape of an eight-year-old girl in India and discussing the lack of accountability of sexual abusers in India for the rape of girls (Ayyub, 2018).

In the case of NSE deepfakes created and distributed to commit IBSA, existing laws against IBSA may not be applicable. The applicability of existing law depends on the scope of the law and its ability to cover technologically manipulated media. In the United States, there is no federal law criminalizing IBSA. While there are no U.S. federal laws, there are, however, various state laws that could be used to prosecute IBSA (see Table 1). Nonetheless, these laws may not apply to AI-manipulated media, such as NSE deepfakes. For this reason, in a couple of U.S. states, laws were amended to criminalize NSE deepfakes. For example, in New York, the law was amended to ensure that it captures NSE deepfakes under N.Y. Penal Code § 245.15 ("Unlawful dissemination or publication of an intimate image"). A person can be charged with § 245.15(1) if the person "with intent to cause harm to the emotional, financial or physical welfare of another person, they intentionally disseminate or publish a still or video image depicting such other person with one or more intimate parts exposed or engaging in sexual conduct with another person, including an image created or altered by digitization, where such person may reasonably be identified from the still or video image itself or from information displayed in connection with the still or video image; and…the actor knew or reasonably should have known that the person depicted did not consent to such dissemination or publication."

There are multiple steps involved in the creation and distribution of NSE deepfakes from the development to the distribution to its hosting on online platforms or via apps. At each stage, the consent of the target of the NSE deepfakes is violated. For adults, consent of the

Table 1 Example of states' laws that could potentially be used for IBSA

Disclosure of intimate visual material (Miss. Code § 97-29-64.1; Texas Penal Code 21.16)
Nonconsensual disclosure/dissemination of private sexual images (Illinois Criminal Code § 11-23.5; Minnesota Criminal Statutes § 617.261; Revised Statutes of Missouri § 573.110; North Dakota Century Code § 12.1-17-07.2; Oklahoma Statutes § 1040.13b; Code of West Virginia § 61-8-28a)
Posting private image for pecuniary gain (Colorado Revised Statutes § 18-7-108)
Prohibition on nude or sexual explicit electronic transmissions (Georgia Code § 16-11-90)
Representation depicting a nude or partially nude person or depicting a person engaging in sexually explicit conduct (Wisconsin Statute § 942.09)
Revenge porn (Maryland Criminal Code § 3-809)
Threatening the nonconsensual dissemination of private sexual images (Revised Statutes of Missouri § 573.112)
Unauthorized dissemination/disclosure of private image (Louisiana R.S. § 14:283.2; Maine Criminal Code 17-A § 511-A; North Carolina General Statutes § 14–190.5A; Ohio Revised Code § 2917.211; Rhode Island General Laws § 11-64-3)
Unlawful/unauthorized distribution/dissemination/disclosure of images/intimate/sensitive image/images depicting nudity or sexual activities/sexual images or recordings (Arizona Revised Statutes § 13-1425; Arkansas Code § 5-26-314; California Penal Code § 647(j)(4); Connecticut General Statutes § 53a-189c; D.C. Code § 20–275; Indiana Code § 35-45-4-8; Kentucky Revised Statutes § 531.120; Michigan Criminal Law § 750.145e; Nebraska Revised Statute § 28-311.08; Nevada Revised Statutes § 200.780; New Mexico Criminal Code § 30–37 A-1; New York § 245.15; Oregon Revised Statutes § 163.472; 18 Pennsylvania Consolidated Statutes § 3131; Utah Code § 76-5b-203; 13 V.S.A. § 2606; Code of Virginia § 18.2-386.2; Washington Revised Code § 9A.86.010; Wyo. Stat. § 6-4-306)
Unlawful exposure (Tennessee Code § 39-17-318)
Video voyeurism (Idaho Code 18-6609(2)(b))

Source: Goldberg, C.A., PLLC. (n.d.). States with Revenge Porn Laws. https://www.cagoldberglaw.com/states-with-revenge-porn-laws/#1613158311027-b1e759b1-acb

person depicted in the NSE deepfake, including the right to revoke consent at any time, should be the only criteria used to determine whether the deepfake created and/or distributed is unlawful, and not the reason why the deepfake was created and/or distributed. When laws include requirements such as "maliciously promot[ing]" (U.S. Florida Statute § 836.13), "knows or should reasonably know would cause a reasonable person to suffer emotional or physical distress or harm" (U.S. Utah Code, § 76-5b-205), "intends to cause distress to that individual" (s. 33 of the UK Criminal Justice and Courts Act of 2015), or "intends to cause …fear, alarm or distress or … is reckless as to whether …[the target] will be caused fear, alarm or distress" (s. 2 Abusive Behavior and Sexual Harm (Scotland) Act of 2016), perpetrators can evade prosecution by claiming that they developed and/or distributed the deepfake without intending to cause harm and for myriad alternative reasons, such as claiming it was meant to be art, meant it to be flattering or fun, to obtain recognition among peers or social recognition, or sexual gratification, among other reasons (for offender motives, excuses, and harms caused, see Henry et al., 2019; Henry & Flynn, 2019). Considering that IBSA is expected to cause harm and understood as causing harm to victims, Australia has laws that do not require prosecutors to show that perpetrators of IBSA intended to cause harm and the IBSA caused harm to victims (e.g., 474.17A of the Criminal Code Act of 1995; for an overview of Australian state, territory and federal law, see RMIT University, Image-Based Abuse Project). Similarly, in the United Kingdom, the Online Safety Act of 2023, added an offense to the Sexual Offences Act of 2003, Section 66B(1), that criminalizes the nonconsensual sharing and the threatening to share an intimate photograph or film without requiring prosecutors to prove that the offender intended to harm the victim. The mere existence of NSE deepfakes without the subject's consent is enough to cause the women depicted in them harm. When laws include provisions that require victims to prove that NSE deepfakes caused them "substantial emotional distress" (see Idaho's enacted House Bill No. 575, which specifically criminalized the nonconsensual disclosure of explicit synthetic media that would cause an "identifiable person substantial emotional distress"), it places an undue and unjustifiable burden on the victim to prove they were harmed. Given the nature of nonconsensual NSE deepfakes, victims should not have to prove that harm was caused.

Cyberbullying and cyberharassment

NSE deepfakes could be used to cyberharass or cyberbully a target. Women who campaigned against NSE deepfakes, after being a target of them, were the targets of even more deepfakes, received threats, and were subjected to cyberharassment and online abuse, for their efforts to remove NSE deepfakes of nonconsenting women (McDermott & Davies, 2022; Plaha & Lee, 2022). A person who uses NSE deepfakes to cyberbully or cyberharass a target could be charged with cyberbullying or cyberharassment laws. In the United Kingdom, Davide Buccheri was convicted of harassment for creating fake nude images of a coworker (not deepfakes; but cheapfakes) and posting them on pornographic websites after she rejected his advances (BBC News, 2018). UK harassment law was used to charge and convict Buccheri for the creation of SE manipulated media. In the United States, state laws on harassment, cyberharassment, sexual harassment or related laws, such as harassment/posting private image for harassment (e.g., Code of Alabama § 13A-6-240; Colorado Revised Statutes § 18-7-107; Iowa Code 708.7; 18 Pa. C.S. § 2709), sexual cyberharassment (Florida Statutes § 784.049), "use or dissemination of visual recording or photographic device without consent and with intent to self-gratify, harass, or embarrass" (South Dakota Codified Laws § 22-21-4), and "distribution of private images with intent to coerce, harass, intimidate or threaten" (Alaska Statutes § 11.61.120), could be used to prosecute NSE deepfakes. In fact, in the U.S., in Pennsylvania, a woman was charged with harassment (18 Pa. C.S.§ 2709) and cyberharassment (18 Pa. C.S. § 2709(a.1)) for creating and distributing deepfakes of high school cheerleaders, including nude deepfakes and deepfakes depicting them smoking and drinking (Lenthang, 2021). The charges relating to the deepfakes were later dropped as insufficient evidence was provided that the media was manipulated with deepfake technology (Harwell, 2021).

NSE deepfakes have also been used to bully targets. Bullying laws could potentially be used to prosecute individuals who use NSE deepfakes to bully others. The ability of the law to apply in these circumstances depends on the scope of the law and if it includes (or at the very least does not restrict the application of the law based on the means used to bully the target). In Australia, in 2023, the eSafety Commissioner reported that children were using CSAM deepfakes (inappropriately described as AI-generated sexually explicit content of children) to bully other students (Long, 2023). In 2023, CSAM deepfakes were also used by a male teenager to cyberbully girl classmates in Spain (Narvali et al., 2023). Specifically, several girls in Spain reported receiving deepfake images to their smartphones. The images were created with a deepfake app that took the clothes off of images where the girls were fully clothed (Llach, 2023). That same year, U.S. girls in a New Jersey high school were the subjects of CSAM deepfakes created by male classmates (Ryan-Mosely, 2023). Moreover, in one school in Pennsylvania in the United States, eighth grade students impersonated their teachers, creating TikToks in their names and likenesses, and

"posted disparaging, lewd, racist and homophobic videos," including videos that implied that their teachers were sexual predators (e.g., "a fake profile…posted a real photo of …[the teacher] at the beach with her husband and their young children", with the following question and answer over the photo: "Do you like to touch kids?" …"Answer: Sí.") (Singer, 2024).

In the United States, concerns were raised that CSAM deepfakes would only be considered cyberharassment and not CSAM, due to the absence of laws criminalizing AI-generated CSAM (Chan & Tenbarge, 2023). This concern is not unfounded as in 2021 a similar concern was raised in New York. That year, a 20-year-old man was arrested for posting NSE deepfakes on a pornography website. This man superimposed the images of several of his female classmates when they were underage onto pornographic images to make it look as if these girls were engaging in sexually explicit acts (McNally, 2023). Along with the CSAM deepfakes he posted, he included victims' names, home addresses, and phone numbers (i.e., doxed them) and encouraged others to contact, threaten, and harass them (CBS New York, 2021). For his crimes, the offender received six months imprisonment and had to register as a sex offender (Gusiff, 2023).

Cyberharassment and cyberbullying laws, like certain IBSA laws, have limitations. The penalties for violations of these laws are not severe, often resulting in short sentences of imprisonment and/or probation. The short penalties for creating and distributing NSE deepfakes do not reflect the harm caused to the targets of deepfakes. These deepfakes can remain online indefinitely and require lifelong actions from victims to remove NSE deepfakes online. Furthermore, youth who develop and/or distribute NSE deepfakes that target children or adults, often face few, if any, consequences, making it difficult to deter this act and repair the harm done. For instance, in the TikTok video mentioned earlier, certain students were temporarily suspended and two students posted an 'apology' video where they stated that what they did was a joke, was blown out of proportion by the teachers, and that they would continue to post content but keep the content private in the future "[be]cause then they [i.e., the teachers and the school] can't do anything" about the videos (Singer, 2024). Schools are limited in what they can do when students post NSE deepfakes outside of school hours (Singer, 2024). Nevertheless, certain schools and school districts in the U.S. have banned the use of cellphones during school hours in part to minimize misuse of the technology (see, for example, Florida, Maine, and Virginia; Singer, 2023).

Sextortion

NSE deepfakes can be used to commit sexual extortion (or sextortion). Sextortion involves threats to share, distribute or otherwise make available intimate information and media unless the demands of the perpetrator(s) are met (e.g., the sharing of sexually explicit images or videos, money, goods, or anything else demanded). The U.S. Federal Bureau of Investigation (2023) warned that NSE deepfakes were being used to extort targets and coerce them into engaging in the perpetrators' demands, such as providing sexually explicit material (e.g., images and/or videos) or providing a payment (e.g., money or gift cards), goods, or services to them. In 2023, sextortion victims reported to the U.S. Internet Crime and Complaint Center that perpetrators threatened to release manipulated media depicting the victim nude or engaging in sexually explicit conduct to others (e.g., family or friends) unless money was paid and/or sexually explicit material was provided to the perpetrators (FBI, 2023).

In other countries, targets were threatened with the release of NSE deepfakes created by offenders using non-sexually explicit media (e.g., available images of the target, images of target taken by the offender, and non-sexually explicit video calls with the target). For example, in one case with multiple victims involving a dating platform known as OKCupid, several victims reported being the targets of nude deepfake images after matching and interacting with someone on the platform (Cabael, 2024). The offenders created nude deepfake images from Telegram video call conversations with victims and subsequently contacted the targets and threatened to release the nude deepfake images they created unless the victims paid money to the offenders (Cabael, 2024).

In the United States and abroad, extortion laws could be used to prosecute NSE deepfakes that are used to commit extortion. In Australia, sextortion is criminalized under extortion laws and state or territorial IBSA laws. Other countries have laws that specifically criminalize sextortion (such as some U.S. states; see, for example, 18 Pa. C.S. § 3133). Recently, in 2024, certain U.S. states proposed and/or enacted laws that criminalize the use of NSE deepfakes for extortion purposes. For example, Idaho enacted House Bill No. 575, which specifically criminalized "threaten[ing] to disclose explicit synthetic media with the intent to obtain money or other valuable consideration from an identifiable person portrayed in whole or in part in the explicit synthetic media."

Violation of privacy

Privacy, a fundamental human right, is enshrined in many constitutions and international human rights instruments. Privacy is polyvalent and enables individuals, among other things, to set and maintain boundaries with others and determine if and under what circumstances access to their information, bodies, property, and residences are warranted. NSE deepfakes could violate laws that prohibit invasions of privacy and the misuse of

private information, including images and videos. However, the conditions under which the protection exists and to what extent it exists varies by jurisdiction. Some jurisdictions outright ban infringements of privacy. In Saudi Arabia, Article 6(1) of the Anti-Cyber Crime Law of 2007, proscribes the "[p]roduction, preparation, transmission, or storage of material impinging on public order, religious values, public morals, and privacy, through the information network or computers." Other jurisdictions ban certain types of invasions of privacy. For example, California's Penal Code 647(j) criminalizes invasions of privacy in the form of secret recordings, where such infringements intrude into one's personal life and space without just cause. In other states in the U.S., NSE deepfakes could potentially be prosecuted with laws that prohibit violations of privacy and/or breach of privacy. However, this depends on the law and whether manipulated media that does not necessarily include the nude body or the body of the target in a sexually explicit position would be covered. For instance, Hawaii's law that proscribes violations of privacy (the Hawaii Revised Statutes § 711-1110.9) would cover NSE deepfakes as it criminalizes the following conduct: "intentional… creat[ion] or disclos[ure] or threaten[ing] to disclose an image or video of a composite fictitious person depicted in the nude…, or engaged in sexual conduct… that includes the recognizable physical characteristics of a known person so that the image or video appears to depict the known person and not a composite fictitious person, with intent to substantially harm the depicted person with respect to that person's health, safety, business, calling, career, education, financial condition, reputation, or personal relationships, or as an act of revenge or retribution." Unfortunately, like other laws targeting IBSA, this law requires victims to not only show that they were harmed by the invasion of privacy caused by the NSE deepfakes, but to also show that they were 'substantially harmed.'

Defamation

NSE deepfakes can be used to defame individuals. Defamatory statements are false and/or misleading statements that can damage an individual's reputation. Certain jurisdictions criminalize defamation. Articles 372 and 373 of United Arab Emirates' Penal Code prohibits "publicity which exposes the victim to public hatred or contempt" and "a false accusation that dishonours or discredits the victim in the public eye." Likewise, in Poland,

> *Whoever imputes to another person, a group of persons, an institution or organisational unit not having the status of a legal person, such conduct, or characteristics that may discredit them in the face of public opinion or result in a loss of confidence necessary for a given position, occupation or type to activity shall be subject to a fine, the penalty of restriction of liberty or the penalty of deprivation of liberty for up to one year. If the perpetrator commits the act specified in § 1 through the mass media shall be subject to a fine, the penalty of restriction of liberty or the penalty of deprivation of liberty for up to [two] years (Article 212, Penal Code).*

In Saudi Arabia, the "[p]roduction, preparation, transmission, or storage of material impinging on public order, religious values, public morals, and privacy, through the information network or computers" is considered a cybercrime (Article 3 of the Anti-Cyber Crime Law of 2007). Furthermore, in India, Sect. 499 of the Penal Code holds that: "Whoever by words either spoken or intended to be read, or by signs or by visible representations, makes or publishes any imputation concerning any person intending to harm, or knowing or having reason to believe that such imputation will harm, the reputation of such person, is said, except in the cases hereinafter excepted, to defame that person." In Japan, perpetrators who sold NSE deepfakes depicting Japanese celebrities (actresses and singers) were arrested for defamation for insulting the honor of the target and damaging their reputation (Ryall, 2020). These deepfakes were posted on a website where users would pay to view the NSE deepfakes. The women depicted in the NSE deepfakes also reported experiencing online abuse (Ryall, 2020).

In the United States, individuals could sue the creator and/or distributor of NSE deepfakes for defamation if factually untrue statements were made about the subject of the deepfake and the deepfake resulted in demonstrated reputational harm. In certain jurisdictions, the target of the deepfake must also show that the person who created and distributed the deepfake intended to cause the target emotional distress (Gieseke, 2020). Litigation, however, is not a reasonable and effective remedy for NSE deepfakes because this legal option unjustly prevents most of the population from obtaining this remedy for economic reasons. To put it simply, these remedies are out of reach for much of the population. What is more, defamation law in the United States does not protect all persons equally. Tiered protection exists against defamation depending on whether a person is a public figure or private citizen.

The NSE deepfake must purport to be real for defamation law to apply (i.e., the manipulated media was intended as a statement of fact). NSE deepfake creators and distributors can add the word 'fake' (or a synonym) in the title of the deepfake, title of the subject or in the message of the email where the deepfake is attached, in a post with a deepfake, or anywhere else to evade liability and use a loophole in the law to avoid experiencing any consequences for creating and/or distributing the NSE deepfake. If NSE deepfakes are distributed with

disclaimers, this diminishes the likelihood that defamation claims will be successful. However, prosecutors and judges should consider that disclaimers can be missed by viewers of deepfakes and edited out by further distributors of the deepfake. Today, with content being consumed quickly but uncritically, the NSE deepfake can spread and be taken on face value as legitimate. Also, if deepfakes are of low quality, making it easy to discern that the video is fake, a defamation claim cannot be brought. Prosecutors and judges need to consider that harm can occur regardless of whether the deepfake is of low or high quality and even if it is identified as a fake.

Intellectual property crime

Intellectual property crime involves "the access, distribution, and/or use of intellectual property without and/or beyond initial authorization and in violation of the rights of the owner or owners of the intellectual property" (UNODC, 2019). An example of an intellectual property crime is copyright infringement. NSE deepfakes could violate copyright laws. However, copyright claims are complicated and will depend on the content depicted (Ray, 2021). The success of copyright infringement claims depends on whether the subject of the NSE deepfakes owns the media content (i.e., took the image or video that was used to create the deepfake and/or has legal rights over that media). The copyright holder of the original content would have to file a copyright claim. The person whose image was superimposed in the video or image cannot file a copyright claim if the person is not the original creator of the content (i.e., the video or image upon which the person's face was superimposed). NSE deepfake creators could bypass copyright claims by claiming fair use of the work for "criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research" (17 U.S.C. § 107). Overall, copyright laws are insufficient to deal with NSE deepfakes. What also complicates copyright claims are 'fair use' claims because a deepfake is not the original content but an outcome of algorithms that splice and combine content to create the deepfake.

In the United States, a copyright holder can bring forth a copyright claim pursuant to the Digital Millennium Copyright Act of 1998. The copyright holder can request to have their copyrighted content removed online using a DMCA takedown. A person can be charged with perjury if they falsely certify that they are the owner or manager of the owner of the copyrighted work. While there have been instances of public figures and celebrities taking down deepfakes, this practice varied by online platform. In 2020, when the deepfake video depicting Nancy Pelosi slurring her words was circulated online, YouTube removed the video as a violation of its community guidelines and policies, but Facebook kept it on the platform with a fact check label as 'false' (Holmes, 2020). Concerns were raised about the efficacy of using copyright takedown notices by non-public figures when reports circulated that Kim Kardashian removed a deepfake depicting her from YouTube, when in fact, the copyright holder, Conde Nast, took down the content as the video that was superimposed was a Vogue interview with Kardashian (Katz, 2019). The copyright owner filed a claim using YouTube ContentID system, which enables copyright holders to take down content that violates their copyrights (Cole, 2019).

Even when DMCA takedowns can be used, an option only available to copyright holders, this measure is, by itself, insufficient to deal with illegal and nonconsensual NSE deepfakes. There are also practical issues that arise when attempts are made to remove these deepfakes online. When the content is taken down on one platform, it can reappear on other platforms. Public figures and celebrities (e.g., female politicians, actresses, musicians, cosplayers, gamers, and streamers), as well as private citizens have been the targets of NSE deepfakes (Cole, 2023; Oppenheim, 2023; Maras & Alexandrou, 2019). The targets of the deepfakes have the painstaking task of identifying content that depicts them on platforms and paying to have the content removed from online platforms by hiring legal teams to take the content off the platform (Oppenheim, 2023). Paying legal professionals to take content off platforms is not an option available to all (only those who can afford it) and does not guarantee that the content will be taken offline. Content posted online could remain there indefinitely as it can be downloaded and reposted online. The victims of the illegal and nonconsensual NSE deepfakes would have to proactively scan sites to identify their deepfakes and take action to ensure it is removed. Neither of these remedies are sufficient and viable solutions for victims. The costs of these actions can be prohibitive and only available to those who have the financial resources and time to pursue them. DMCA takedowns, therefore, are not an adequate remedy for NSE deepfakes (except the few who can utilize this option) and do not guarantee permanent removal of NSE deepfakes. The question that follows is: are there other measures that could be used to deter and minimize the spread of illegal and nonconsensual NSE deepfakes?

The herculean task of countering deepfakes

The countering of deepfakes is an Augean task, which is a term derived from Greek mythology[3] that describes an extremely difficult task that requires intervention to improve adverse and detrimental situations with

[3] In the *Labors of Hercules*, King Eurystheus ordered Hercules to perform labors, one of which was the cleaning of the Augean stables with divine livestock, which had never been cleaned, in one day. Given the task before him, Hercules came up with an ingenious solution, which did not involve him

deleterious effects. The countering of deepfakes requires a multifaceted response that spans legal, policy, regulatory, ethical, and technological measures, while recognizing the outright inimical impact of NSE deepfakes on the subjects' human rights. While the recommendations included in the next sections represent what we view as the most optimal solution for countering NSE deepfakes, we will utilize the forthcoming subsections to convey the merits, drawbacks, and current societal and legal constraints of these approaches. It is our aim that this realistic level setting will enable researchers and policymakers to make informed determinations regarding the feasibility and deficiencies in proposed measures to address illicit and nonconsensual NSE deepfakes.

Mind the legislative gaps: updating laws and providing guidance on the applicability of laws

Overall, variations exist in laws on CSAM deepfakes and nonconsensual NSE deepfakes, including the extent to which they cover, if at all, CSAM deepfakes and nonconsensual NSE deepfakes, making protections against these deepfakes and remedies afforded to victims dependent on jurisdiction. For these reasons, existing laws should be harmonized across jurisdictions, to the extent possible, and amended to fill these gaps.

Certain new laws (and amendments to existing laws) that were recently enacted or proposed do not adequately address these gaps and deal with CSAM deepfakes and nonconsensual NSE deepfakes. For example, in the U.S., laws that were recently enacted to deal with CSAM deepfakes in certain U.S. states do not cover CSAM deepfakes that do not depict an identifiable minor (i.e., a real child) (e.g., Idaho House Bill No. 485). U.S. state laws should criminalize CSAM deepfakes even if they do not depict real children (e.g., California's proposed 2024 Assembly Bill No. 1831); this practice is aligned with U.S. federal law that criminalizes this material as child obscenity. Existing laws apply to very specific situations, offering few, if any, legal remedies for targets of nonconsensual NSE deepfakes. For example, copyright laws could apply if a person is the copyright owner of the media (e.g., video; not the person depicted in it unless they own the copyright to the video). IBSA laws, if present, can only be applied to specific situations and if intent to harm is shown (with few exceptions, e.g., Australia), which offenders can bypass in many jurisdictions with laws that contain this provision by claiming other reasons for the development and distribution of nonconsensual NSE deepfakes. Cyberbullying and cyberharassment laws can also only be applied to specific situations involving nonconsensual NSE deepfakes. For instance, in the UK, a poet and broadcaster, Helen Mort, was the target of nonconsensual NSE deepfakes, where an abuser uploaded non-intimate images of her online, which were taken from her private Facebook account, and "encouraged other users to edit her face into violent pornographic photos" (Hao, 2021). Because the abuser engaged in actions that were not covered by existing law (e.g., not using the victim's real name and not personally creating the photos), the victim had no legal recourse against the abuser for the harassment (Hao, 2021).

Furthermore, laws that are enacted should provide adequate remedies for those targeted by NSE deepfakes. Currently, U.S. federal law does not provide targets of nonconsensual NSE deepfakes with effective remedies. Specifically, the *Disrupt Explicit Forged Images and Non-Consensual Edits* (DEFIANCE) Act of 2024 proposed in the U.S. as a law that helps those targeted by NSE deepfakes (Ocasio-Cortez, 2024) is short-sighted and does not provide an effective remedy for all victims. This law provides a civil right of action, which can only be pursued by those who have significant financial resources to sustain the cost of civil lawsuits, making this 'right' only available to those who can afford it. A better measure would ensure that legal remedies are accessible and available to all, regardless of financial means. An example of this is a different U.S. federal law that was proposed in response to nonconsensual NSE deepfakes, the *Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks* (TAKE IT DOWN) Act, which bans the publication of nonconsensual intimate imagery (NCII), including nonconsensual NSE deepfakes, outright, including its use as a tool for sextortion, and "require[s] social media and similar websites to have in place procedures to remove…[NCII] content upon notification from a victim" (U.S. Senate Committee on Commerce, Science & Transportation, 2024). This law, considers the primary role of the platforms in hosting and distributing nonconsensual NSE deepfakes, and places a legal obligation on these platforms to remove this content - a similar obligation already exists for copyrighted material.

Finally, guidelines should be implemented on how existing laws can be used to adequately address deepfakes and the harm they cause. These guidelines can assist criminal justice agents in interpreting existing laws and their applicability to NSE deepfakes. These guidelines would, therefore, help with the enforcement of the laws.

ISP guidelines, responsibilities, and liability

Online intermediaries (or Internet intermediaries), such as search engines, social media platforms, and internet service providers that supply Internet access and related services, provide "access to, host, transmit and index content, products and services originated by third parties on

physically cleaning the stables (i.e., diverting rivers to clean the stables with the flood of water).

the Internet or provide Internet-based services to third parties" (Perset, 2010, p. 9 cited in UNODC, 2021, p. 15). Internet intermediaries where deepfakes are hosted, shared, and distributed, play an essential role in countering deepfakes. Internet intermediaries have taken different approaches to NSE deepfakes – some have taken them off their platforms and servers, others have kept them but added disclaimers, and others have done nothing at all and/or encouraged the uploading of NSE deepfakes to their platforms (e.g., numerous websites were created to house NSE deepfakes either exclusively or partially; see Burgess, 2023).

Internet intermediaries are not adequately dealing with NSE deepfakes. A 2024 U.S. news report revealed that search engines (e.g., Google, Bing and DuckDuckGo) place NSE deepfakes and associated tools to develop them as top results (Gabrielli, 2024). This practice makes it easier for individuals to access NSE deepfakes and the tools used to create them. These search engines have the ability to suppress (or downgrade) the results, which they have done for other websites (e.g., those that contain mugshots; see Segal, 2013), but have not taken this action against NSE deepfakes. What is more, the response of these search engines and other Internet intermediaries is to place the onus to identify and report NSE deepfakes to Internet intermediaries on the target of the NSE deepfakes (or others who are willing to report it), with no guarantee of removal as there is no legal obligation to remove them. Internet intermediaries also predominantly do not proactively detect, remove, and prevent NSE deepfakes and tools that develop them from being posted and distributed on their platforms and/or appearing in top search results.

Even though Internet intermediaries have policies that prohibit illegal content and other cyber-enabled crime, these cybercrimes persist. Platforms, such as Facebook/Meta, Twitter/X, and TikTok, among others, have terms of use/rules/community guidelines that prohibit users from abusing, bullying and harassing others, posting and distributing objectionable content (adult nudity and sexual activity), distributing materials with nonconsensual nudity, engaging in adult sexual exploitation, committing child sexual exploitation and abuse, engaging in gender-based violence, and spreading misinformation (i.e., incorrect information). Various forms of synthetic and manipulated media are also banned by certain online platforms. For example, in its community guidelines, TikTok bans synthetic and manipulated media that contains the likeness of a private figure. They also prohibit the use of synthetic and manipulated media that depict public figures in a manner that violates community guidelines, such as to harass these public figures. These policies and community guidelines, however, are not consistently enforced (Forrester, 2023) and rely on the targets of NSE deepfakes (or others who come across it and want to report it) to inform the platforms of the existence of prohibited content and conduct (see reporting policies and practices of social media platforms).

Internet intermediaries are not universally required to consider and take action to ensure the safety and security of users. In certain jurisdictions, Internet intermediaries are regulated and required to implement measures to protect the safety, security, and rights of users of the platform. For instance, the European Union passed the Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (a.k.a. the Digital Services Act), which "regulates online intermediaries and platforms such as marketplaces, social networks, content-sharing platforms, app stores, and online travel and accommodation platforms. Its main goal is to prevent illegal and harmful activities online and the spread of disinformation [i.e., purposely false information]. It ensures user safety, protects fundamental rights, and creates a fair and open online platform environment" (European Commission, n.d.). The Act requires Internet intermediaries to protect children, ensure the rights of users are respected, and provide users of platforms with tools to report and request the removal of illegal products and illegal content from the platforms.

Nevertheless, in other jurisdictions, Internet intermediaries are not held liable for content hosted on their platform. In the United States, Internet intermediaries, pursuant to Sect. 230 of the Communications Decency Act (CDA) of 1996, have immunity over content hosted on their platforms, with a few exceptions. Under 47 USC 230(c)(1), "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." The law includes limited exceptions to Internet intermediary liability immunity; the most recent of which pertains to sex trafficking offenses under the Allow States and Victims to Fight Online Sex Trafficking Act of 2017 (FOSTA) (see Brannon & Holmes, 2024, for an overview of Sect. 230). Section 230 of the CDA also does not provide Internet intermediaries with immunity from liability for copyright violations. Accordingly, Internet intermediaries must remove hosted content that violates copyright law. In the U.S., the proposed TAKE IT DOWN Act of 2024, if enacted, would create legal liability for nonconsensual NSE deepfakes, and would require Internet intermediaries to take down reported content within 48 h of the receipt of a victim's report. Legal liability would incentivize the proactive enforcement of policies and community guidelines and removal of content that violates their policies and guidelines (Smith & van Alstyne, 2021), such as nonconsensual NSE deepfakes.

Even though some countries hold Internet intermediaries liable for violating legal obligations and the law, most countries do not hold them liable for failing to report illegal content, like CSAM. Moreover, while countries do not require Internet intermediaries to proactively monitor content for illegal content, some countries do require that Internet intermediaries immediately contact law enforcement (or other designated agencies or organizations) if they are notified of specific illegal content and/or encounter specific illegal content, while others require Internet intermediaries to take some action if illegal content is identified (ICMEC, 2023).

Platforms must prioritize obtaining proper consent from depicted individuals when uploading media (audio, video, images). Additionally, detection software should be accessible to both platforms and users to identify deepfake media being described as authentic content. To provide this capability, it is necessary to conduct an evaluation of the existing detection software and create a toolkit that enables users to easily verify the legitimacy of media by uploading content or providing media URLs. In the case where a URL is provided and the media is confirmed to be a deepfake, the platform should receive automatic notification and be required to acknowledge that the user has identified the media as a deepfake, as well as provided proof of consent from the individual depicted in the deepfake. In many cases this can be an automated process with two systems communicating and an individual only needing to get involved if consent is being contested by an individual claiming to be depicted in the media. The open-source nature of deepfake technology presents challenges in registering each deepfake media created during the production process. However, companies that aim to provide these services legally can be subject to regulation and compelled to register all media produced by their products. While there are scholars who suggest educating the general public about deepfakes (Achyut, 2023; Prasanna Shashikant Shinde, 2024), we hold a different perspective due to the arduous nature of correctly identifying high-quality deepfakes. In lieu of this, we believe that public campaigns would provide greater utility if they disseminated information about detection software that can identify deepfakes, while directing users to software that is free to access and provided and maintained by platforms and government entities.

Ethical considerations for deepfake technology developers

Deepfake technology developers should consider the ethical implications of the technology they create and the training datasets they use. For instance, apps that enable users of the app to remove clothing from people in media, such as images, to create nude or semi-nude images, are predominantly designed to enable the development of NSE deepfakes. The only legitimate uses of the app are its use to remove clothing from media that depicts the user of the app or another consenting individual. The uses of the app, however, have not been limited to these applications alone. While some of these apps are promoted as fun and creative ways for users to express themselves, other apps in their marketing explicitly state the singular purpose of their apps (e.g., "See anybody naked for free," Undress.app; "realization of an original AI algorithm for generating nudes from photos of clothed women," DeepNudeNow.com).

The ethical implications of manipulated media have been considered by industries, such as the entertainment and marketing industries. Movie studios and advertising companies have been selective in their use of deepfakes. For example, a few instances of the use of deepfakes include Disney's *The Book of Boba Fett*, Bruce Willis' Megafon advertisement, Tom Hanks' role in the film *Here*, members of the LGBTQ community interview for *Welcome to Chechnya*, and the documentary *Gerry Anderson: A Life Uncharted* (Lees, 2024). Documentary makers have been one of the early adopters of deepfakes to tell the stories and provide a 'true' portrayal of individuals who are deceased or are unable to recreate significant moments. The companies involved in the development of these deepfakes perceive their work as ethical, achieved through the acquisition of consent and input from the subject's family members and estate, along with the appointment of an Ethics officer within their organizations (Lees, 2024). Responsible officers have been mandated in the U.S. for securities with the Sarbanes–Oxley Act of 2002 and in the EU required to enforce compliance with General Data Protection Regulation (GDPR). A legally-mandated responsible officer may be beneficial to regulate deepfake companies and ensure the content generated using their software is created in a responsible and ethical manner. The use of ethical deepfakes in documentaries shows that deepfakes can be used ethically since documentaries are a form of nonfiction truth telling and are still able to achieve this goal by using a type of media that by definition is considered manipulated. Finally, as individuals and companies decide whether they should create a particular deepfake, the guiding questions should be: (1) Is producing this deepfake necessary to convey a message or achieve a particular goal? (2) Does the deepfake cause individual, organizational, group, community, or societal harm? (3) Are there mitigation measures in place to prevent or mitigate this harm?

The developer of DeepNude, an app which removes the clothes from women in images, withdrew the app after "receiving widespread attention" (Kastrenakes, 2019). This app was designed to objectify women and its adverse impact on women was reasonably foreseeable,

despite claims to the contrary, making the developer at least partially responsible for the harm caused to women through the use of the app (Johnson & Diakopoulos, 2021). Johnson and Diakopoulos (2021) call for developers to "design tools and techniques that limit the possibility of harmful or dangerous use" (p. 34). One way this could be accomplished is to mandate 'ethics by design'[4] in the development stage of AI tools. Ethical considerations of AI technology include its reasonable and proper use, informed consent of those depicted in the deepfakes, the protection of the rights, such as privacy, of the person depicted in the deepfakes, the ability to identify the deepfake and the person who made the deepfake, and the legitimacy of what is being depicted in the deepfake (Li & Wan, 2023). The European Commission (2021) had identified several ethical principles that should be preserved and protected by AI technology. The ethical principles that should guide the development of AI tools that enable the development of NSE deepfakes are "respect for human agency," which encompasses human dignity and personal autonomy that underlie fundamental human rights, and contribution to "individual, social, and environmental well-being" (European Commission, 2021, p. 5). AI tools that respect human agency do not "subordinate…, coerce…, deceive…, manipulate…, objectif[y] or dehumanize" individuals (European Commission, 2021, p. 6) and AI tools that contribute positively to individual and social well-being do not cause individuals harm (European Commission, 2021, p. 8–9). NSE deepfakes neither respect human agency nor contribute positively to individual and social well-being. Instead, NSE deepfakes objectify and dehumanize women and violate their "right to digital self-representation," which "establishes… that others may not manipulate digital data that represent people's image and voice, as markers of the self, in hyper-realistic footage that presents them in ways to which they would object" (deRuiter, 2021, p. 1327–1328) and "in ways that disregard their will and go against their sense of self" (deRuiter, 2021, p. 1327). The right to digital self-representation thus necessitates informed, affirmative consent of the subject to be represented in the deepfakes (which can be revoked at any time) and the "protection against the manipulation of hyper-realistic digital representations of …[individuals] image[s] and voice[s]" as "a fundamental moral right" (deRuiter, 2021, p. 1328).

Digital signatures: Watermarks and digital fingerprinting

Digital signatures, such as watermarks and other forms of digital fingerprinting, have been proposed as general countermeasures for deepfakes. Digital fingerprints of NSE deepfakes could be created to readily identify deepfakes and prevent this content from being uploaded online. Watermarking is a digital fingerprinting option used to uniquely identify each deepfake produced by a particular software for a specific user (Chen et al., 2012; Emmanuel & Kankanhalli, 2006; Maani et al., 2008; Nikolaidis & Pitas, 2006). In China, to be considered legal, manipulated media must include identifying markers, like watermarks. While it would be preferable to have all manipulated media registered, we acknowledge it is unfeasible. What we can achieve is digital signatures and watermarks for deepfakes created for legitimate use, deepfake software distributed by legitimate marketplaces, or deepfakes created using the processing power of companies, such as Amazon and Microsoft. First, deepfake software hosted on platforms like App Store, Google Play, and GitHub should be required by these platforms to watermark and provide content IDs for the media created using deepfake software. Platforms like GitHub should also be required to implement this system since ML and AI researchers have utilized GitHub to host their software. In addition, services, such as Amazon Web Services (AWS), should require the identification of deepfake software utilizing their service, and all manipulated media created using this service should be watermarked and have content IDs. These platforms should either independently or collectively create a library which adds a watermark and content ID to all outputs created using deepfake software stored on their platform or utilizing their processing power. Alternatively, other methods could be used. For example, other digital fingerprinting methods could use media content, such as a video's content, to create a digital fingerprint, making the digital fingerprint dependent on the content and not on embedding a watermark or digital ID into the content (Chen et al., 2012; Maani et al., 2008). This method, unlike watermarks, produces a more reliable and consistent identifier if appropriately implemented. It would require the use of a technique that uses unique characteristics of the video, which can be processed quickly and makes a digital fingerprint file.[5]

Content sharing platforms should require their users who upload media to the platform to indicate whether the media is manipulated. If the content uploaded is NCII, including NSE deepfakes, consent from the person depicted in NCII must be obtained to upload it to the website. Next, all content sharing platforms should be required to check the deepfake content ID database to ensure the media uploaded by a user is not a deepfake or is a deepfake that has not been labeled appropriately (i.e., as a form of manipulated media) and/or does not adhere to content policies. Material not indicated as a deepfake, but which have a deepfake watermark or a content ID,

[4] 'By design' signals the consideration of what precedes the phrase 'by design' at the outset (at the design stage).

[5] Fingerprinting videos is more difficult than fingerprinting images.

should be prevented from being uploaded to the platform. When manipulated media is uploaded to the platform, it should have a clear label identifying the media as manipulated. Finally, while this method would cover most of the manipulated media generated, illegal and nonconsensual NSE deepfakes would require additional tools to identify them as a deepfake (these methods are discussed in the following section, sect. Utilizing research meant to improve deepfakes to spot flaws in manipulated media). Many of the tools and methods identified in the following section have been implemented in Meta's deepfake detection toolkit (Yin, 2021), Google (Kite-Powell, 2023), and Intel (Intel newsroom, 2022). Finally, one crucial element still missing is the standardization of banned content across content sharing platforms. Manipulated media banned from one platform should be removed from other platforms and prevented from being uploaded to other platforms unless the ban has been rescinded.

However, we must acknowledge that digital signatures are only effective for platforms and applications that agree to have them integrated into their software. The existence of open-source deepfake software allows for content creation that bypasses consent requirements and application-level checks, placing the burden of preventing the dissemination of this content on platforms, which is a role that some platforms may be reluctant to take without legislative requirements. For this reason, laws are needed that create legal liabilities for Internet intermediaries.

Utilizing research meant to improve deepfakes to spot flaws in manipulated media

The improvement and detection of deepfakes is commonly referred to as a "cat-and-mouse game" between cybersecurity and criminals (Kerner & Risse, 2021). This analogy does not capture the fact that many of the improvements in the quality of deepfakes have come from researchers from higher education institutions in recent years. Several papers, when presenting new methods for generating deepfakes, start by identifying a current imperfection in deepfakes and either hypothesize or test a technique capable of fixing the imperfection, which provides individuals who are using deepfakes for nefarious purposes ways to improve the quality of future content (Averbuch-Elor et al., 2017; He et al., 2019; Jung et al., 2020; Suwajanakorn et al., 2017). While these research articles have hypothesized ways to correct imperfections in deepfakes, these methods have not become standard, and not everyone has access to software that has implemented these methods or the technical expertise to implement them themselves. The gap between known imperfections and standardized implementation of fixes presents an opportunity to flag and remove deepfakes containing explicit content depicting an individual who has not provided consent to the use of their image in that manner.

Many imperfections commonly found in deepfakes are typically present in facial features, including the eyes, mouth, and teeth. Researchers found that to improve the quality of mouth movement, rather than transferring the mouth from one video to the next, they could synthesize the mouth, improving the depiction quality (Suwajanakorn et al., 2017). While this method was computationally expensive and produced short clips, a new method was quickly developed, which only required text to be edited and reduced the mismatch between the mouth movement and audio (Fried et al., 2019). Other researchers have developed programs to change specific characteristics of the face while preserving other characteristics (He et al., 2019). While these techniques have been created to improve the quality of deepfakes, the flaws outlined can be implemented into detection systems on platforms that allow users to share content.

While several flaws have been identified in research seeking to improve the quality of deepfakes, a small but meaningful number of papers have taken the time to identify characteristics and digital artifacts present within deepfakes. Jung et al. (2020) developed DeepVision, which detected anomalies in eye blinks with 87.5% accuracy when tested on different videos. Digital forensics has also provided several methods for verifying whether videos have been tampered with. The SIFT-based forensics method was developed to determine the brand and type of device used to take the image and the modifications made (Amerini et al., 2011). Facebook has indicated they have developed newer methods to uncover similar digital artifacts in deepfakes (Diaz, 2021). An increase in fake identification documents like passport, driver license, and health insurance cards have been facilitated by deepfakes created using morphing techniques. Many of these documents or the templates to create these documents are currently cheaply available on many darknet marketplaces. Demorphing has been proposed as a technique to identify deepfakes created by transforming e-documents. Kramer et al. (2019) present a method for reducing the number of documents capable of fooling automatic border control (ABC) systems and humans at a country's port of entry using a demorphing technique, which showed significant success over both humans and ABC systems when identifying transformed e-documents. Finally, other researchers have discussed keeping track of deepfake content using smart contracts. The contracts would create a hash of the deepfake content along with storing additional information about the content and would allow platforms to check content before allowing users to upload deepfakes to their platforms (Hasan & Salah, 2019). Other researchers have developed more specific frameworks to determine the

authenticity of a video by collecting and analyzing other media about a particular event to determine the likelihood of its authenticity (Amerini et al., 2017).

Identifying imperfections in deepfake media and integrating detection systems in platforms that allow them to detect deepfakes has taken on renewed importance. This becomes apparent with the availability of deepfake nude generator software and deepfake models becoming available for purchase on darknet marketplaces starting in 2023.

Conclusion

Deepfakes and their harmful impacts have been discussed by academics, news media, politicians, and practitioners since 2017; and yet viable solutions have not been universally implemented. Instead, discussions continue to center on the harmful impacts of NSE deepfakes with insufficient forward movement and actions taken to stop, slow, or minimize the spread of these deepfakes, deter creators and distributors of these deepfakes without the subject's consent, hold NSE deepfakes creators and distributors accountable, and mandate actions to be taken by Internet intermediaries to prevent or at the very least minimize the availability and distribution of illicit and nonconsensual NSE deepfakes. What remains additionally problematic is that many of the techniques currently implemented to identify and keep track of deepfakes are based on techniques and methods predating deepfakes, meaning these safeguards could have been deployed with deepfake technology, and not as it was done, well after its deployment.

Outlawing all deepfakes is unrealistic and unfeasible. However, existing laws need to be reviewed, amended, and updated to ensure that nonconsensual NSE deepfakes, as well as NSE deepfakes that contain illegal content and/or are used to commit other forms of cyber-enabled crime, are prohibited. These modifications are needed to close legal loopholes to ensure that nonconsensual NSE deepfakes, as well as NSE deepfakes that depict illegal content (i.e., CSAM) and are used to commit technology-facilitated gender-based violence, including image-based sexual abuse, cyberharassment, cyberbullying, sextortion, invasion of privacy and defamation, among other crimes, are prohibited. Following a review of existing laws and identification of the modifications needed, guidelines should be developed to assist prosecutors and judges in understanding the applicability of existing laws to illegal NSE deepfakes, as well as nonconsensual NSE deepfakes and illegal uses of them.

Amendments to laws and guidelines on the applicability of law are not all that is needed to counter deepfakes. Existing legal measures, particularly civil lawsuits, are not available equally to all individuals due to the exorbitant financial costs of litigation. This process is also ineffective in the long term as content that is uploaded tends to remain online indefinitely due to downloads and reposting of original content. In view of that, proactive responses are needed to prevent illegal and nonconsensual NSE deepfakes from being uploaded online at the outset.

Technology can be used to identify and prevent the upload of NSE deepfakes and to minimize their distribution. Currently, intervention occurs after NSE deepfakes are posted and users report these deepfakes to Internet intermediaries. This after-the-fact action (i.e., waiting until after NSE deepfakes are uploaded) occurs too late to effectively deal with NSE deepfakes and mitigate harm to deepfake subjects. Given that content is consumed quickly but uncritically, NSE deepfakes can spread and be taken by viewers on face value as legitimate even if they are identified as a fake, thus harming the subjects of the deepfakes regardless of the intent of the deepfake creator and/or distributor. NSE deepfakes should thus be banned outright on online platforms, unless express, affirmative, and informed consent is obtained from the person depicted in the NSE deepfake, which can be revoked at any time. Internet intermediaries should thus be legally liable for and be required to remove illicit and nonconsensual NSE deepfakes from their platforms.

Ultimately, existing countermeasures do not adequately deal with illegal and nonconsensual NSE deepfakes, and do not sufficiently focus on the actual individuals responsible for the NSE deepfakes and their proliferation - the offenders who create and distribute them, the deepfake developers who create technology specifically designed to create NSE deepfakes of unsuspecting and nonconsenting targets, and the Internet intermediaries that do not take proactive measures to ban NSE deepfakes on their platforms. This lack of focus on those ultimately responsible for illegal and nonconsensual deepfakes is seen even in well intentioned public service announcements. For example, in 2023, Germany's Deutsche Telekom shared a deepfake of a 9-year-old as a warning against 'sharenting' - the practices of parents sharing images of their children, which can be used to feed AI technology to create deepfakes (Johnson, 2023). Restricting online actions and participation because of the actions of others has a chilling effect on the public. The victims are not the problem, the offenders are, and the lack of adequate remedies for victims and the infringement of their human rights online.

Acknowledgements
Not applicable.

Author contributions
Marie-Helen Maras, Conceptualization; Investigation; Resources; Writing - Original Draft; Writing - Review & Editing; Project Administration; Supervision. Kenji Logie, Investigation; Resources; Writing - Original Draft; Writing - Review & Editing.

Funding
This work was supported by the Center for Cybercrime Studies, John Jay College of Criminal Justice, City University of New York.

Data availability
Not applicable.

Declarations

Competing interests
The authors declare no conflicts of interest.

Received: 13 January 2024 / Accepted: 5 September 2024
Published online: 18 October 2024

References

Achyut, T. S. (2023). DeepFake Deception: A comprehensive analysis of DeepFake Technology and its effects on Ethics. *Politics and Society International Journal of Scientific Research in Engineering and Management*, *07*(09). https://doi.org/10.55041/IJSREM25653

Adjer, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf September 2019.

Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2011). A SIFT-based forensic method for copy–move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, *6*(3), 1099–1110. https://doi.org/10.1109/TIFS.2011.2129512

Amerini, I., Becarelli, R., Brancati, F., Caldelli, R., Giunta, G., & Itria, M. L. (2017). Media trustworthiness verification and event assessment through an integrated framework: A case-study. *Multimedia Tools and Applications*, *76*(5), 7197–7212. https://doi.org/10.1007/s11042-016-3303-8

Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, *27*(4), 1–13. https://doi.org/10.1093/jcmc/zmac008

Averbuch-Elor, H., Cohen-Or, D., Kopf, J., & Cohen, M. F. (2017). Bringing portraits to life. *ACM Transactions on Graphics*, *36*(6), 1–13. https://doi.org/10.1145/3130800.3130818

Ayyub, R. (2018). I was the victim of a deepfake porn plot intended to silence me. *Huffington Post*, November 21, 2018. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

Bae, G., & Yeung, J. (2023). South Korea has jailed a man for using AI to create sexual images of children in a first for country's courts. *CNN*, September 27, 2023. https://www.cnn.com/2023/09/27/asia/south-korea-child-abuse-ai-sentenced-intl-hnk/index.html

BBC News (2018). City worker 'posted fake porn photos of colleague' April 18, 2018.

Brannon, V. C., & Holmes, E. N. (2024). Section 230: An Overview. Congressional Research Service, R46751. https://crsreports.congress.gov/product/pdf/R/R46751

Burgess, M. (2023). Deepfake Porn Is Out of Control. Wired, October 16, 2023. https://www.wired.co.uk/article/deepfake-porn-is-out-of-control

Cabael, K. (2024). Scammers using deepfake nude images to demand money from victims in Singapore. *Straits Times*, June 29, 2024. https://www.straitstimes.com/singapore/courts-crime/scammers-using-deepfake-nude-images-to-demand-money-from-victims-in-singapore

CBS New York (2021). Long Island man faces felony charges for posting 'deepfake' images of former classmates on porn site. *CBS News New York*, December 16, 2021. https://www.cbsnews.com/newyork/news/long-island-man-faces-felony-charges-for-posting-deepfake-images-of-former-classmates-on-porn-site/

Chan, M., & Tenbarge, K. (2023). For teen girls victimized by 'deepfake' nude photos, there are few, if any, pathways to recourse in most states. NBC News, November 23, 2023. https://www.nbcnews.com/news/us-news/little-recourse-teens-girls-victimized-ai-deepfake-nudes-rcna126399

Chen, Y. H., Huang, H. C., & Wang, S. Y. (2012). Video scrambling and fingerprinting for digital right protection. *2012 International Symposium on Computer, Consumer and Control*, 471–474. https://doi.org/10.1109/IS3C.2012.125

Cole, S. (2019). The Kim Kardashian deepfake shows copyright claims are not the answer. *Vice*, June 19, 2019. https://www.vice.com/en/article/j5wngd/kim-kardashian-deepfake-mark-zuckerberg-facebook-youtube

Cole, S. (2023). Deepfake porn creator deletes internet presence after tearful 'Atrioc' Apology. *Vice*, January 31, 2023. https://www.vice.com/en/article/jgp7ky/atrioc-deepfake-porn-apology

Confederation of European Data Protection Organisations (2023). Generative AI.

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*, 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

DeepFaker App (2022, May 1). Privacy Policy. DeepFaker App. https://deepfaker.app/privacy_policy/

Del Valle, G. (2024). Wisconsin man arrested for allegedly creating AI-generated child sexual abuse material. *The Verge*, May 21, 2024. https://www.theverge.com/2024/5/21/24161965/ai-csam-instagram-stable-diffusion-arrest

Diaz, J. (2021). Facebook researchers say they can detect deepfakes and where they came from. *NPR*. https://www.npr.org/2021/06/17/1007472092/facebook-researchers-say-they-can-detect-deepfakes-and-where-they-came-from

Doffman, Z. (2019). Chinese deepfake app ZAO goes viral, privacy of millions 'at risk.' *Forbes*, September 2, 2019. https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-face-app-like-privacy-storm/?sh=de8573584700

Emmanuel, S., & Kankanhalli, M. S. (2006). Mask-based fingerprinting scheme for digital video broadcasting. *Multimedia Tools and Applications*, *31*(2), 145–170. https://doi.org/10.1007/s11042-006-0041-3

European Commission (2021). Ethics By Design and Ethics of Use Approaches for Artificial Intelligence (V. 1). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

European Commission (n.d.). The Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

FBI (2023). Malicious actors manipulating photos and videos to create explicit content and sextortion schemes. Public service announcement, I-060523-PSA (June 5, 2023). https://www.ic3.gov/Media/Y2023/PSA230605

Forrester (2023). Fundamental Problems On Social Media Platforms. *Forbes*, February 28, 2023. https://www.forbes.com/sites/forrester/2023/02/28/fundamental-problems-on-social-media-platforms/

Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., Genova, K., Jin, Z., Theobalt, C., & Agrawala, M. (2019). Text-based editing of talking-head video. *ACM Transactions on Graphics*, *38*(4), 1–14. https://doi.org/10.1145/3306346.3323028

Gabrielli, V. (2024). Google and Bing put nonconsensual deepfake porn at the top of some search results.

Gieseke, A. P. (2020). The new weapon of choice: Law's current inability to properly address deepfake pornography. *Vanderbilt Law Review*, *9*(5), 1479–1515.

Gusiff, C. (2023). Patrick Carey sentenced to 6 months for deepfaking images of young women, putting them on porn site. *CBS New York*, April 19, 2023. https://www.cbsnews.com/newyork/news/patrick-carey-sentenced-to-6-months-for-deepfaking-images-of-young-women-putting-them-on-porn-site/

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology Behavior and Social Networking*, *24*(3), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

Hao, K. (2021). Deepfake porn is ruining women's lives. Now the law may finally ban it. *MIT Technology Review*, February 12, 2021. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

Harwell, D. (2021). Remember the 'deepfake cheerleader mom'? Prosecutors now admit they can't prove fake-video claims. *The Washington Post*, May 14, 2021. https://www.washingtonpost.com/technology/2021/05/14/deepfake-cheer-mom-claims-dropped/

Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *Ieee Access : Practical Innovations, Open Solutions*, *7*, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, *28*(11), 5464–5478. https://doi.org/10.1109/TIP.2019.2916751

Henry, N., & Flynn, A. (2019). Image-based sexual abuse: Online distribution channels and illicit communities of support. *Violence against Women*, *25*, 1932–1955.

Henry, N., Flynn, A., & Powell, A. (2019). Image-based sexual abuse: Victims and perpetrators. *Trends & Issues in Crime and Criminal Justice*, No. 572 (March 2019). Australian Institute of Criminology, https://www.aic.gov.au/sites/default/files/2020-05/imagebased_sexual_abuse_victims_and_perpetrators.pdf

Holmes, A. (2020). A doctored video that makes Nancy Pelosi appear drunk went viral on Facebook — again. *Business Insider*, August 3, 2020. https://www.businessinsider.com/nancy-pelosi-facebook-declines-to-remove-doctored-viral-video-2020-8

Intel Introduces Real-Time Deepfake Detector. Intel newsroom, & Intel (2022). https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html

International Center for Missing and Exploited Children (2023). *Child Sexual Abuse Material: Model Legislation & Global Review* (10th Edition).

Johnson, M. (2023). Deutsche Telekom creates AI deepfake 'girl' in new identity abuse campaign. *Mediashotz*, July 5, 2023. https://mediashotz.co.uk/telekom-launches-disturbing-ai-deepfake-identity-abuse-campaign/

Johnson, D. G., & Diakopoulos, N. (2021). Computing Ethics: What to do about deepfakes. *Communications of the ACM*, *64*(3), 33–35.

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes detection using human eye blinking pattern. *Ieee Access : Practical Innovations, Open Solutions*, *8*, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Kastrenakis, J. (2019). Controversial deepfake app DeepNude shuts down hours after being exposed. *The Verge*, June 27, 2019. https://www.theverge.com/2019/6/27/18761496/deepnude-shuts-down-deepfake-nude-ai-app-women

Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

Kinnard, M. (2023). 'A race against time:' Prosecutors in all 50 states urge rules to fight child-abuse images in AI. September 5, 2023. https://fortune.com/2023/09/05/child-abuse-images-ai-state-prosecutors/

Kite-Powell, J. (2023, August 31). *Google* launches tool that detects ai images in effort to curb deepfakes. *Forbes*. https://www.forbes.com/sites/jenniferhicks/2023/08/31/google-launches-tool-that-detects-ai-images-in-effort-to-curb-deepfakes/

Lees, D. (2024). Deepfakes in documentary film production: Images of deception in the representation of the real. *Studies in Documentary Film*, *18*(2), 108–129. https://doi.org/10.1080/17503280.2023.2284680

Lenthang, M. (2021). Cheerleader's mom created deepfake videos to allegedly harass her daughter's rivals. *ABC News*, March 13, 2021. https://abcnews.go.com/US/cheerleaders-mom-created-deepfake-videos-allegedly-harass-daughters/story?id=76437596

Li, M., & Wan, Y. (2023). Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research*, *33*(5), 1750–1773.

Llach, L. (2023). Naked deepfake images of teenage girls shock Spanish town: But is it an AI crime? *Euronews*, September 24, 2023. https://www.euronews.com/next/2023/09/24/spanish-teens-received-deepfake-ai-nudes-of-themselves-but-is-it-a-crime

Long, C. (2023). First reports of children using AI to bully their peers using sexually explicit generated images, eSafety commissioner says. *ABC News*, August 15, 2023. https://www.abc.net.au/news/2023-08-16/esafety-commisioner-warns-ai-safety-must-improve/102733628

Maani, E., Tsaftaris, S. A., & Katsaggelos, A. K. (2008). Local feature extraction for video copy detection in a database. *2008 15th IEEE International Conference on Image Processing*, 1716–1719. https://doi.org/10.1109/ICIP.2008.4712105

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of Artificial Intelligence and in the wake of Deepfake Videos. *International Journal of Evidence and Proof*, *23*(3), 255–262.

Maras, M. H., & Logie, K. (2021). Understanding what it really takes to control your data: A critical evaluation of Faceapp. *Journal of Internet Law*, *24*(8), 1–18.

McDermott, S., & Davies, J. (2022). Deepfaked: 'They put my face on a porn video.' *BBC News*, October 21, 2022. https://www.bbc.co.uk/news/uk-62821117

McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'revenge porn': The continuum of image-based sexual abuse. *Feminist Legal Studies*, *25*(1), 25–46.

Mcnally, K. (2023). Seaford man sentenced to 6 months in jail for posting 'deepfake' sexual images of women. *News 12 Long Island*, April 18, 2023. https://longisland.news12.com/seaford-man-to-be-sentenced-for-posting-deepfake-sexual-images-of-women

Mezzofiore, G. (2018). No, Emma Gonzalez did not tear up a photo of the Constitution. *CNN*, March 26, 2018. https://www.cnn.com/2018/03/26/us/emma-gonzalez-photo-doctored-trnd/index.html

Narvali, A. M., Skorburg, J. A., & Goldenburg, M. J. (2023). Cyberbullying girls with pornographic deepfakes is a form of misogyny. *The Conversation*, November 28, 2023. https://theconversation.com/cyberbullying-girls-with-pornographic-deepfakes-is-a-form-of-misogyny-217182

Nikolaidis, N., & Pitas, I. (2006). Image and video fingerprinting for digital rights management of multimedia data. *2006 International Symposium on Intelligent Signal Processing and Communications*, 801–807. https://doi.org/10.1109/ISPACS.2006.364920

O'Brien, W., & Maras, M. H. (2024). Technology-Facilitated Coercive Control: Response, Redress, Risk, and Reform. *International Review of Law, Computers and Technology*, published online January 12, 2024.

Ocasio-Cortez, A. (2024). Rep. Ocasio-Cortez Leads Bipartisan, Bicameral Introduction of DEFIANCE Act to Combat Use of Non-Consensual, Sexually-Explicit Deepfake Media. Press release, March 7, 2024. https://ocasio-cortez.house.gov/media/press-releases/rep-ocasio-cortez-leads-bipartisan-bicameral-introduction-defiance-act-combat

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of deepfake pornography. *Ethics and Information Technology*, *22*, 133–140.

Oppenheim, M. (2023). 'I had to pay to get deepfake porn removed of me': Presenter reveals the dark side of being a female gamer. *Independent*, April 10, 2023. https://www.independent.co.uk/news/uk/home-news/sunpi-porn-deepfake-gamers-youtube-b2315465.html

Perset, K. (2010). The economic and social role of internet intermediaries, OECD Digital Economy Papers, No. 171 (OECD Publishing, Paris). https://doi.org/10.1787/5kmh79zzs8vb-en

Plaha, M., & Lee, J. (2022). Sharing pornographic deepfakes to be illegal in England and Wales. *BBC News*, November 25, 2022. https://www.bbc.com/news/technology-63669711

Prasanna Shashikant Shinde. (2024). Deepfakes and their impact on Society. *International Journal of Advanced Research in Science Communication and Technology*, 283–290. https://doi.org/10.48175/IJARSCT-15749

Reuters (2020). Fact check: Drunk Nancy Pelosi video is manipulated. https://www.reuters.com/article/idUSKCN24Z2B1/

RMIT University (n.d.). Laws in Australia. The Image-Based Abuse Project. https://www.imagebasedabuse.com/the-laws-in-australia/

Ryall, J. (2020). Celebrity deepfake porn cases in Japan point to rise in sex-related cybercrime. *South China Morning Post*, November 20, 2020.https://www.scmp.com/week-asia/lifestyle-culture/article/3110748/deepfake-porn-cases-japan-point-rise-sex-related

Ryan-Mosley, T. (2023). A high school's deepfake porn scandal is pushing US lawmakers into action. *MIT Technology Review*, December 1, 2023. https://www.technologyreview.com/2023/12/01/1084164/deepfake-porn-scandal-pushing-us-lawmakers/

Sadiq, M. (2019). Real v fake: debunking the 'drunk' Nancy Pelosi footage - video. *Guardian*, May 24, 2019. https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video

Segal, D. (2013). Mug-Shot Websites, Retreating or Adapting: One site stops charging fees for removing arrest photos. But another resumes the practice. *New York Times*, November 10, 2013. https://www.nytimes.com/2013/11/10/your-money/mug-shot-websites-retreating-or-adapting.html

U.S. Senate Committee on Commerce, Science & Transportation (2024). Sen. Cruz Leads Colleagues in Unveiling Landmark Bill to Protect Victims of Deepfake Revenge Porn. Press Releases, June 18, 2024. https://www.commerce.senate.gov/2024/6/sen-cruz-leads-colleagues-in-unveiling-landmark-bill-to-protect-victims-of-deepfake-revenge-porn

Singer, N. (2023). This Florida School District Banned Cellphones. Here's What Happened. *New York Times*, October 31, 2023. https://www.nytimes.com/2023/10/31/technology/florida-school-cellphone-tiktok-ban.html

Singer, N. (2024). Students Target Teachers in Group TikTok Attack, Shaking Their School. *The New York Times*, July 6, 2024. https://www.nytimes.com/2024/07/06/technology/tiktok-fake-teachers-pennsylvania.html

Smith, M. D., & van Alstyne, M. (2021). It's Time to Update Sect. 230. *Harvard Buisness Review*, August 12, 2021. https://hbr.org/2021/08/its-time-to-update-section-230

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, *36*(4), 1–13. https://doi.org/10.1145/3072959.3073640

Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI) Europol's European Cybercrime Centre (EC3). (2020). Malicious Uses and Abuses of Artificial Intelligence. https://www.europol.europa.eu/cms/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf

Turk, V. (2019). Deepfakes are already breaking democracy. Just ask any woman. *Wired*, November 18, 2019. https://www.wired.co.uk/article/deepfakes-pornography

United States of America v. Steven Anderegg. (2024b). Indictment, 24-CR-50-JDP (U.S. District Court, Western District of Wisconsin, May 15, 2024). https://www.justice.gov/opa/media/1352606/dl?inline

United States of America v. Steven Anderegg (2024a). Government's Brief in Support of Detention, 24-CR-50-JDP (U.S. District Court, Western District of Wisconsin, May 20, 2024), https://www.justice.gov/opa/media/1352611/dl?inline

UNODC (2021). Issue Paper Policymaking and the role of online intermediaries in preventing and combating Illicit trafficking (United Nations Publishing, Vienna). https://sherloc.unodc.org/cld/uploads/pdf/Online_intermediaries_eBook.pdf

UNODC (2019). Intellectual property: What it is. Module 11: Cyber-Enabled Intellectual Property Crime. https://sherloc.unodc.org/cld/en/education/tertiary/cybercrime/module-11/key-issues/intellectual-property-what-it-is.html

UNODC (2022). Digest of Cyber Organized Crime, 2nd edition. https://www.unodc.org/documents/organized-crime/tools_and_publications/21-05344_eBook.pdf

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

Vincent, M. (2023). Bossier man jailed for child porn also state's first to face new deepfake law. *Fox 8 Live*, December 27, 2023. https://www.fox8live.com/2023/12/27/bossier-man-jailed-child-porn-also-states-first-face-new-deepfake-law/

Wang, S., & Kim, S. (2022). Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, *134*, 107305. https://doi.org/10.1016/j.chb.2022.107305

Yin, X. (2021). Detecting the models behind deepfakes. *Meta Newsroom*, June 16, 2021. https://about.fb.com/news/2021/06/detecting-the-models-behind-deepfakes/

Publisher's note

Marie-Helen Maras is a tenured Full Professor and the Director of the Center for Cybercrime Studies at John Jay College of Criminal Justice. Dr. Maras' academic background and research cover transnational crime, particularly cybercrime, cyberlaw and cyberliberties, and the impact of digital technology. She has a DPhil in Law and an MSc and MPhil in Criminology and Criminal Justice from the University of Oxford. Dr. Maras serves as a subject matter expert and consultant on cybercrime and cyber organized crime for UNODC. She is the author of numerous peer-reviewed academic journal articles and books, including *Real Criminology* (Oxford University Press), *Cybercriminology* (Oxford University Press), and *Computer Forensics: Cybercriminals, Laws, and Evidence* (Jones and Bartlett), among other books.

Kenji Logie is a Ph.D. student in the Criminal Justice program at John Jay College of Criminal Justice and a research associate at the John Jay College of Criminal Justice Center for Cybercrime Studies. His research interests and publications include cybercrime, darknet forums, darknet marketplaces, and digital forensics.