4. Privacy, Pornography, and AI: A Multidisciplinary Exploration of Deepfakes

Introduction

This paper reflects on the experience of a friend, pseudonymously named Emma, whose privacy rights were infringed upon by AI through the creation of a deepfake video of her. This incident, having received Emma's consent for discussion, serves as the starting point for a broader exploration of ethical concerns in Big Data, including privacy, security, and rights (Kuc-Czarnecka & Olczyk, 2020). The reflection delves into the issue from legal, philosophical, and data science perspectives, adopting a multidisciplinary approach deemed necessary for the complex research area of ethics in Big Data (Kuc-Czarnecka & Olczyk, 2020). The first part of this paper looks at previous work by relevant scholars to set the stage for my proposals, starting with outlining Emma's experience and introducing the concept of deepfakes. It then delves into the societal impacts of deepfake pornography and reviews the current protective measures and legal landscapes. The second part, which is primarily my original contribution, attempts to find a balance between regulation and freedom of expression, proposes new legal and regulatory frameworks, and underscores the vital role of data scientists and AI researchers in addressing the challenges posed by harmful deepfakes.

Anatomy of a Deepfake: The Technology Behind Emma's Experience

Six months ago, Emma, 21, discovered a viral video of herself singing and dancing on TikTok, which was confusing given that she was not the person in the footage. She was disturbed to learn that someone had created a deepfake of her. In spite of the video's non-explicit nature, she felt her privacy was violated as a result of its unauthorised creation and dissemination using her likeness—sourced from public social media content. Despite her efforts, the platform's investigation did not favour the video's removal as the tag #deepfake was used in the post’s caption (TikTok Creator Portal, 2023). Lacking clear legal options in South Carolina for taking down the online video, Emma had no choice but to accept the video's online permanence. This incident served as a wake-up call for our friend group, forcing us to examine our

digital footprints as young women, fearing the worst-case scenario of deepfake exploitation: deepfake pornography.

The term "deepfake" originated from a Reddit user named "[u/]deepfakes" in 2017 to refer to images or recordings that have been effectively altered and edited to portray someone as doing or saying something they did not do or say (Cole, 2017). Deepfakes are created using machine learning and a specific deep learning application called Generative Adversarial Networks (GAN), where two self-supervised algorithms learn from each other (Gieseke, 2020). Through numerous training cycles, these algorithms become capable of generating hyper-realistic synthetic images that are difficult to distinguish from real ones with the human eye (Chesney and Citron, 2019). In GANs, one model (the discriminative algorithm) classifies input data, while the other (the generative model) creates data indistinguishable from the original dataset (Gieseke, 2020). The advancement of deepfake technology is deeply intertwined with the quantity of available data to train algorithms, often sourced from social media platforms (Cole, 2017; Chesney and Citron, 2019; Gieseke, 2020).

The political repercussions of deepfakes have frequently been the focus of academic and media debate, particularly the erosion of faith in government and its implications for democracy (Gieseke, 2020). Furthermore, the macro-level epistemic consequences of deepfakes in disseminating disinformation are becoming increasingly concerned. This idea is similar to Chesney and Citron's (2018) concept of the "liar's dividend," in which actual wrongdoing is discounted as just another deepfake. However, as Kugler and Pace (2021) point out, deepfakes can inflict individual injury to their subjects' dignity and emotional well-being in addition to the greater social impact. Focusing on their second dimension of harm and D'Ignazio and Klein's (2020) concern with the lack of neutrality in the age of big data, this paper integrates feminist philosophy and feminist legal theory to scrutinise the consequences of enabling and permitting the production and distribution of deepfakes online. This approach provides a multifaceted perspective on the broader social and legal ramifications of AI and deepfake technology, recognising the fact that data, including the data used in creating and training deepfake algorithms, are influenced by social, political and historical contexts (D'Ignazio and Klein, 2020).

Deepfake Pornography: Emma's Worst-Case Scenario

96% of online deepfake videos are pornographic in nature, with 99% of pornographic deepfakes targeting women (Kugler and Pace, 2021, Mania, 2022). This alarming trend was highlighted when a Reddit user, u/deepfakes, posted a manipulated video featuring a famous actress, spurring the creation of a subreddit dedicated to such content (Cole, 2017). These technologies lower the barrier for creating fake pornographic videos using images sourced from social media, making any woman who has shared images of herself online (clothed or not) a potential victim (Gieske, 2020).

Deepfake pornography, as exemplified by over 100,000 cases on Telegram the span of a few months (Hsu, 2023), signifies a disturbing evolution of revenge pornography, as perpetrators can create pornographic material of any woman with an online presence to threaten or hurt them with (Eshete, 2021). The creation of pornographic deepfakes in this context can be understood to inflict individual injury to their subjects' dignity and emotional well-being as discussed by Kugler and Pace (2021). Deepfake pornography has already impacted many individuals' (mainly women's) reputations and job prospects, as around 80% of employers conduct online searches that yield these manipulated results and just one explicit deepfake can prevent a woman from being hired due to societal norms (Gieseke, 2020; Eshete, 2021). The following section examines the current measures in place (or lack thereof) to protect people from such individual injury by primarily focusing on the legal and regulatory landscape in the United States. The focus on the United States in this analysis stems not from an Americentric viewpoint but from two primary reasons: firstly, it is the jurisdiction where Emma encountered difficulties in seeking justice, and secondly, it is a significant hub for deepfake pornography, hosting 41% of websites and online platforms with such content (Mania, 2022).

Fortress or fence? Examining the Effectiveness of Current Safeguards

Self-policing

As illustrated by Emma's experience, one of the first and most accessible lines of defence against the damaging effects of deepfakes are the online platforms that host them. Platforms often engage in self-policing, where they independently set and enforce guidelines to manage and regulate content (Wood and Sanders, 2020).

While Reddit banned the r/deepfakes subreddit for violating community standards, the damage was already done with deepfake pornography spreading to other platforms (Gieseke, 2020). Platforms like X (formerly Twitter) have taken proactive steps by flagging and removing deepfakes, but others, including Facebook and TikTok, have been less stringent, especially regarding non-explicit content (Wood and Sanders, 2020; Pangburn, 2019). This is likely because, from a market perspective, platforms lack a strong incentive to ban deepfakes. As highlighted by Pangburn (2019) there exists a complete “misalignment of incentives," where social media platforms are driven to promote deepfakes due to the high user engagement they generate.

This slightly differs when it comes to deepfake pornography where some platforms such as those run by Meta have existing policies against nudity (Cole, 2017), which prevent the sharing of explicit content, however, this approach is not always effective nor uniformly applied across all platforms. Regardless of some policies flagging deepfakes, platforms and websites which allow explicit content like Pornhub, Gfycat and X still host nonconsensual deepfake pornography due to the technological challenge of detecting them (Gieseke, 2020; Cole, 2018). In addition to the absence of a market incentive to self-police, Section 230 of the Communications Decency Act, as highlighted by the Electronic Frontier Foundation, offers legal immunity to platforms for user-generated content to protect freedom of expression (Brown, 2019; Wood and Sanders, 2020). This means platforms do not face the fear of lawsuits based on such posts and rely on this protection for their continued existence and minimal intervention when it comes to deepfakes and AI-generated content (Brown, 2019; Wood and Sanders, 2020). The limitations of self-policing by online platforms introduce us to the ongoing challenge of aligning market, technological, legal, and social forces in regulating big data (Boyd and Crawford, 2012) and deepfakes to be further explored in the remainder of this section.

Existing Laws

In the United States, where the majority of explicit and non-explicit deepfakes are hosted, the legal landscape surrounding deepfakes and the more extreme case of deepfake pornography is fragmented. While states like New York and California have specific laws banning deepfake pornography, there is no unified national approach (Hsu, 2023). While it could be argued that defamation and copyright laws could be

used to address this issue, I argue that such laws in their current state are not enough. Defamation laws, including libel, which encompasses defamatory content on the internet, struggle to address the complexities of deepfakes. One major hurdle is the requirement for victims to prove the intent of emotional distress by the creators, which is often challenging (Gieseke., 2020). Deepfake creators may not anticipate their victims discovering the content, making intent hard to establish. Furthermore, even if victims claim copyright over their photos used in deepfakes, the transformative nature of these modifications complicates the issue, adding to the difficulty and cost of legal recourse (Gieseke., 2018; Noreen et al., 2018). This suggests that without existing precedent, victims of malicious deepfakes remain unprotected.

The most relevant Supreme Court case, Ashcroft v. Free Speech Coalition (2002), which considered the validity of the Child Pornography Prevention Act of 1996 in cases involving virtual child pornography, shifted the balance between freedom of speech and obscenity (Gieseke, 2020). While the court recognised the inherent moral blameworthiness of child pornography, the Child Pornography Prevention Act was nonetheless struck on the grounds that it restricted freedom of speech (Gieseke, 2020). The court also made an important distinction between virtual computer-generated depictions and actual harm, suggesting that the Miller framework which allowed the government to restrict obscenity under the First Amendment did not apply (Gieseke, 2020). This is likely because the pornography in question was characterised as a creative speech act such as political cartooning and deepfakes have been defended by applying the same logic. The distinction made is one that I will argue ought to be re-examined in the context of deepfakes particularly deepfake pornography, which, although simulated, can cause real emotional, physical, and financial harm due to the depiction of actual people.

Proposed Legislation

The Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 (“DEEPFAKES” Accountability Act) proposed in the 116th Congress in 2019, would require mandatory watermarks for deepfakes fulfilling a role akin to TikTok’s AI content label (Wood and Sanders, 2020). I argue that this legislative measure has several shortcomings. The first, as outlined by Brown (2019) is that those weaponising deepfakes are likely to ignore the

legislation and share their AI creations anonymously. The second is as illustrated by Emma's case, victims of deepfakes may still feel that their privacy has been violated and even be harmed in spite of the deepfake being labelled as such. This sentiment was echoed by respondents in a study carried out by Kugler and Pace (2021), where a majority of respondents still viewed deepfakes as morally blameworthy and harmful in spite of their clear labelling as fictional. The third, and primary focus of this discussion is the fact that the legislation has yet to be enacted due to concerns over freedom of speech and expression (Hsu, 2023). To elaborate, while the bill's proponent described the bill as a 'protective measure' distinguishing it from China's similar law which she described as a 'control mechanism', it has yet to be accepted due to free speech concerns (Hsu, 2023). The proponent's distinction is telling of two things: the first being the fact that the ways in which we develop, interact with and regulate AI technology can be politically motivated; and the second being the fact that free speech concerns, in this case those related to the First Amendment, often act as a barrier to such regulation in Western democracies (Brown, 2019). In the following sections, I aim to balance effective AI regulation with free speech in order to protect other rights.

Balancing Regulation with Freedom of Expression

Balancing ethical regulation with the imperative of free speech is challenging yet achievable. It is achievable by adopting the multidisciplinary approach proposed by Kuc-Czarnecka and Olczyk (2020) to gauge big data phenomena from a technical and philosophical perspective as an understanding of the social and ethical dimensions of AI can help formulate effective legislation and guidelines. As outlined above, while they protect freedom of expression and speech the First Amendment and Section 230 create barriers which make it difficult to protect the victims of deepfakes from deepfake technology.

The First Amendment ensures that Congress cannot pass laws which infringe on one's freedom of speech, which includes the right to express ideas without interference through the media and is viewed as critical for democracy (Post, 1990). John Stuart Mill's Harm Principle, an exception within this framework, contends that speech should be unrestricted unless it causes harm to others (Mill, 1859). This principle aligns with the U.S. Supreme Court's stance, which allows restricting speech likely to incite harmful unlawful acts (Langton, 1993). Under this framework

speech (including deepfakes) cannot be prohibited for being false or causing offense, however, it can be prohibited for causing harm. Thus, rather than viewing deepfakes as a form of creative expression as was done during the Ashcroft v. Free Speech Coalition case, I suggest we ought to instead treat harmful deepfakes as a harmful speech act as this could justify regulation under the harm principle. By acknowledging the harm caused by deepfakes, such regulation aligns with the broader interests of freedom and democracy, balancing ethical concerns with First Amendment challenges.

I argue that deepfakes, particularly pornographic ones, can be seen as speech acts, echoing feminist philosophers such as Langton (1993) and Mackinnon's (1992) work on the ethics of pornography and censorship. Deepfakes in this context are more than mere artificially created representations; they represent a complex form of speech act that profoundly affects women's presence in public discourse. These artificial representations exacerbate the credibility and authority gaps between genders, reinforcing stereotypes and discrediting women's legitimacy. The fear of misuse for harassment or defamation further silences women, leading to a phenomenon known as "illocutionary disablement," where women's speech is overshadowed by manipulated narratives (Langton, 1993). This unique form of harm, lacks effective counter-speech remedies, as seen illustrated by the limitations of mandatory watermarks, making it a distinct category of harmful expression which not only encourages physical and psychological harm to a group but effectively silences them by restricting their ability to express themselves freely and be heard. Based on the illocutionary disablement deepfakes cause to their victims an argument can be made that restricting them actually promotes people's freedom of speech and equality.

I suggest that deepfakes can also be seen to cause epistemic injustices in non-political contexts in cases where non-influential individuals are targeted. This is because due to their realistic nature, deepfakes contribute to the generation of "dirty data," which perpetuates a cycle of epistemic injustice (Richardson et al., 2019). Dirty data can be understood as data that does not accurately reflect reality but instead mirrors existing biases (Richardson et al., 2019).

For example, in the case of deepfake pornography, the generation of such content results in skewed data that reinforces gender biases. This process creates a self-perpetuating cycle of epistemic injustice, where the distorted portrayal of women leads to further entrenched discrimination and misrepresentation in society. This

cycle is not only a technological issue but also a social and political one, where flawed data generation and usage continually reinforce and exacerbate systemic biases.

Here we might argue that the issue may not be with deepfake technologies and algorithms but with how they have been used, suggesting that biased or unequal outcomes may not be indicative of an inherently flawed or biased technology. To this I turn to look at technologies like DeepNude, which debase female images while failing to function similarly with male images, underscoring the gendered bias inherent in this technology (Gieske, 2020). I argue that the gender-specific targeting by technologies like DeepNude is not just a biased outcome, but a deliberate design choice, reflecting a deeper, systemic bias in AI development. This technology's training to specifically target women reveals an intentional embedding of gender discrimination, going beyond mere algorithmic bias to represent a calculated decision in AI programming. This aspect raises significant ethical questions about the responsibility of developers in perpetuating gender biases through AI and deepfake technologies while also showcasing a clear intent to harm and further stigmatise a specific group of people, something akin to a harmful speech act such as hate speech which instigates further harm i.e. hate crimes (Langton, 1993). The impact of this has been documented by Kugler and Pace (2011) who found nonconsensual online postings of women and sexual minorities can further lead to their stigmatisation.

Before moving on further it is important to stress that while DeepNude and pornographic deepfakes are specific examples that may not necessarily be representative of all deepfake technologies, they serve as critical lenses for examining and understanding these AI outcomes. Drawing on D’Ignazio and Klein (2020), any meaningful analysis should consider the "knowledge infrastructure" from which these technologies emerge. However, the predominance of pornographic deepfakes, especially those targeting women, as outlined before implies that this analysis encapsulates the overarching biases and the harms prevalent in these technologies. It is also worth recognising the fact that not all deepfakes are harmful, even pornographic ones. For example, rather than engaging in sexual activity, actors in the adult film industry could consent to having their likeness used in pornographic deepfakes, potentially minimising risks of exploitation. This issue will be addressed further in the following section.

Building a Fortress: A New Legal and Regulatory Framework

To effectively regulate harmful deepfakes, including deepfake pornography, a multifaceted regulatory approach could be implemented. One potential solution involves amending Section 230 (and similar regulations in other nation-states) to allow victims to sue platforms for failing to remove deepfakes or for inadequate moderation, thus incentivising platforms to strengthen their moderation efforts. This would improve the state of platform self-policing and allow victims to take action in cases where self-policing fails. However, it is worth noting that merely giving victims the opportunity to take legal action against platform may not be the most effective solution as suing large corporations is likely to be costly and time consuming.

I suggest that it may be better to turn to independent agencies such as those that promote consumer protection. In the American context, regulation could be jointly administered by the Federal Trade Commission (FTC) and the Federal Communications Commission (FCC), as both have relevant authority. The FTC, with its mandate to protect consumers against unfair practices, could serve as a reporting point for deepfake violations, especially through its Division of Privacy and Identity Protection (Gieseke, 2020). Furthermore, expanding the FCC's role to oversee social media and online video platforms in the same way that it regulates broadcast media might increase oversight. To ensure accessibility for all victims, victims could file online complaints with such regulators, launching an intra-agency adjudicatory process. This procedure would allow victims to file legal action against platforms or persons responsible for deepfakes. It is important to note that the suggested framework could and should be expanded beyond the United States. Collaboration with international regulatory authorities and adaptation of the FTC and FCC model to diverse legal and cultural situations would be required. Setting worldwide norms and exchanging best practices would promote a coordinated response to the cross-border nature of online material and deepfakes. Such standards could additionally push platforms to engage in more effective self-policing by potentially adopting common guidelines allowing for a more unified approach, potentially leading to some version of Noreen et al.'s (2018) vision of sectoral codes of conduct or agreements. Such international cooperation is vital, given the international reach of deepfake technology and its impacts, ensuring a comprehensive and effective approach to mitigating its harms.

Leveraging AI to Fight Deepfakes at the Source

I suggest that data scientists and AI researchers have the potential to play a greater role in combating the proliferation of harmful deepfakes than lawmakers or regulatory bodies. Recognising their responsibility, given they created the virus, they are uniquely positioned to develop the vaccine. This proactive approach is essential, as pointed out by Wood and Sanders (2020), especially since the damage from deepfakes can occur well before detection and removal. The harm caused by deepfakes persists even after the victim is compensated, as the content may remain online and continue to be redistributed. This underlines the need for effective detection mechanisms. Facebook appear to have been among the first to recognise this, creating Deepfake Detection Challenge, announced in September 2019, incentivising AI researchers to find solutions to this problem (Mania, 2021). This initiative represents a form of 'crowd competition', where a large number of participants, often from diverse backgrounds, compete to solve a problem, harnessing collective intelligence and creativity (Matzler, 2020). As emphasised several times throughout this paper multidisciplinary is necessary in solving unorthodox problems such as the ones posed by deepfakes, and it may be worth exploring other options such as crowd collaboration to reach a solution. Creating effective detection tools is of paramount importance, especially when we compare the ease of creating deepfakes to detecting them (Wood and Sanders, 2020), which can be described as yet another gap between ethical regulation and technological advancement (Kuc-Czarnecka & Olczyk, 2020). Such tools could pre-screen content across various platforms, like Twitter and Facebook.

The addition of filters to original images to prevent their use in deepfakes is a promising approach. Startups such as Truepic and Deeptrace are pioneering in this area with some success (Wood and Sanders, 2020). I propose taking this a step further by integrating such filters into all cameras and having the option to add them to all existing photos in people’s camera rolls and social media feeds. Adopting such a strategy where only consensual deepfakes are generated from future images could be more effective than an outright ban on the technology. This approach not only ensures ethical use but also avoids driving the practice underground or hindering the potential positive outcomes that deepfake technology might offer in various AI applications. For example, in the adult entertainment industry, deepfakes could be

used positively by employing an individual's consented likeness instead of engaging them in activities they are uncomfortable with. Such a use necessitates prior consent, achievable through deliberately taken or shared unfiltered photos specifically for creating these deepfakes. This method respects individual choice and autonomy, providing a safe and ethical alternative within the industry.

Conclusion

In conclusion, by reflecting on Emma's experience, this paper sheds light on the intricate dance between ethics, law, and technology in the era of deepfakes. It illuminates the profound societal and psychological harms, particularly for women, stemming from deepfake pornography and exposes the glaring inadequacy of current legal frameworks. Beyond critique, this work offers a roadmap for navigating this complex terrain, advocating for targeted legislative reforms that prioritize victim protection and empower data scientists and AI researchers through novel mechanisms like crowd competition and collaboration. This multi-pronged approach, integrating ethical, legal, and technological perspectives, represents not just a theoretical ideal but a practical necessity to mitigate the dangers of deepfakes and pave the way for a more responsible and equitable technological future.

References

Boyd, D. and Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), pp.662-679.

Brown, N.I., 2019. Deepfakes Are Frightening, But So Is Congress' Rush to Regulate Them. Slate, [online] 15 July. Available at: [URL] [Accessed 26 December 2023].

Chesney, R. and Citron, D.K., 2018. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. SSRN Electronic Journal. Available at: https://doi.org/10.2139/ssrn.3213954 [Accessed 26 December 2023].

Cole, S., 2017. AI-Assisted Fake Porn Is Here and We're All Fucked. VICE: Motherboard, [online] 11 Dec. Available at: https://www.vice.com/enus/article/gydydm/gal-gadotfake-ai-porn [Accessed 26 December 2023].

Cole, S., 2018. Pornhub Is Banning AI-Generated Fake Porn Videos, Says They're Nonconsensual. VICE: Motherboard, [online] 6 Feb. Available at: https://www.vice.com/en_us/article/zmwvdw/pornhub-bans-deepfakes [Accessed 26 December 2023].

D'Ignazio, C. and Klein, L.F., 2020. Data Feminism. Cambridge: MIT Press. Available at: http://ebookcentral.proquest.com/lib/warw/detail.action?docID=6120950 [Accessed 26 December 2023].

Eshete, B., 2021. Making machine learning trustworthy. Science, 373(6556), pp.743-744.

Gieseke, A.P., 2020. "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. Vanderbilt Law Review, 73, p.1479.

Hsu, T., 2023. As Deepfakes Flourish, Countries Struggle With Response. The New York Times.

Kuc-Czarnecka, M. and Olczyk, M., 2020. How ethics combine with big data: a bibliometric analysis. Humanities and Social Sciences Communications, 7(1), p.137. Available at: https://doi.org/10.1057/s41599-020-00638-0 [Accessed 26 December 2023].

Kugler, M.B. and Pace, C., 2021. Deepfake Privacy: Attitudes and Regulation. SSRN Electronic Journal. Available at: https://doi.org/10.2139/ssrn.3781968 [Accessed 26 December 2023].

Langton, R., 1993. Speech Acts and Unspeakable Acts. Philosophy and Public Affairs, 22(4), pp.293-330.

Mackinnon, C., 1992. Pornography, Civil Rights and Speech. In: Catherine Itzin (ed.) Pornography: Women, Violence and Civil Liberties. Oxford: Oxford University Press.

Mania, K., 2022. Legal protection of revenge and deepfake porn victims in the European Union: findings from a comparative legal study. Trauma, Violence, & Abuse, p.15248380221143772.

Matzler, K., 2020. Crowd innovation: The philosopher's stone, a silver bullet, or pandora's box?. NIM Marketing Intelligence Review, 12(1), pp.10-17.

Mill, J.S., 1975 [1859]. On Liberty. In: Three Essays. Oxford: Oxford University Press.

Nooren, P. et al., 2018. Should We Regulate Digital Platforms? A New Framework for Evaluating Policy Options. Policy & Internet, 10(3), pp.264-301. Available at: https://doi.org/10.1002/poi3.177 [Accessed 26 December 2023].

Pangburn, D.J., 2019. You've Been Warned: Full Body Deepfakes are the Next Step in AI-Based Human Mimicry. Fast Company, [online] 21 September.

Post, R.C., 1990. Racist speech, democracy, and the First Amendment. William & Mary Law Review, 32, p.267.

Richardson, et al., 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice.

TikTok Creator Portal, 2023. AI-Generated Content Label. TikTok Creator Portal. Available at: https://www.tiktok.com/creators/creator-portal/en-us/community-guidelines-and-safety/ai-generated-content-label/ [Accessed 30 December 2023].

Wood, J. and Sanders, N., 2020. Dealing with "Deepfakes": How Synthetic Media Will Distort Reality, Corrupt Data, and Impact Forecasts. Foresight: The International Journal of Applied Forecasting, (59), pp.32-37.