Behind the Screens: Gendered and Generational Divides in Understanding Deepfake Violence

A Qualitative Mixed Methods Study into Deepfake Perceptions and Impacts

Student Name: Anja Ellwood

Student Number: 740726

Supervisor: João Gonçalves, PhD

Master Media Studies - Digitalisation, Surveillance & Societies

Erasmus School of History, Culture and Communication

Erasmus University Rotterdam

Master's Thesis

June 2025

Word Count: 20705

Behind the Screens: Gendered and Generational Divides in Understanding Deepfake Violence

ABSTRACT

In 2024, all eyes turned to South Korea as it became the first victim of an epidemic of AI-generated deepfakes that disproportionately targeted women and minors. This technological threat, now surfacing worldwide, involves the non-consensual creation and distribution of hyper-realistic imagery, with current approaches to regulating this technology highlighting a societal unpreparedness and disconnect in understanding its true impact. This thesis thus explores the intricate issue of gendered digital violence, exploring how different groups perceive these harms, and the unexpected dual role of minors as both victims and perpetrators within this. Grounded in a virtual feminist theoretical framework, this multi-method qualitative research utilises focus groups and critical discourse analysis, with findings uncovering a 'Digital Violation Discrepancy' suggesting that the lack of women's perspectives, in AI development and regulation shapes understandings of deepfake harms. This discrepancy stems from a compounded issue: AI tools, instilled with patriarchal biases, birth an exploitative harm rooted in consent violation, further exacerbated by the crime's sui generis anonymity affordance that disrupts traditional legal and judicial proceedings relying on traceable evidence. These complexities were seen to be less understood by the male perspective, mirroring a critical gap in regulation and development measures that reflects this underrepresentation of women in these spheres. Consequently, deepfake creation's primary consent violation remains inadequately addressed in regulation, reflecting news representation where platform accountability is lacking, and fabricated harms, such as AI-generated child sexual abuse material, are normalised due to a misunderstanding of digital native behaviours. To address this, the research advocates for a foundational paradigm shift that prioritises women's experiences across all phases of AI development, ethical deliberation, and regulatory oversight, spanning AI governance and development reform, strengthening societal and educational responses, and encouraging international cooperation to harmonise legal frameworks. Ultimately, this thesis stresses that achieving an authentically equitable and secure digital future, safe from the uniquely gendered harms of deepfakes, requires challenging existing power structures within technology, ensuring AI empowers rather than exploits, particularly for women and minors.

KEYWORDS: *Artificial Intelligence, Deepfakes, Gendered Violence, Digital Ethics, Victim-Centred, Minors*

Abbreviations:

AI - Artificial Intelligence

CSAM - Child Sexual Abuse Material

NCII - Non-Consensual Intimate Images

AIFA - AI Facilitated Abuse

NSFW - Not Safe For Work

GANs - Generative Adversarial Networks

VAEs - Variational Auto-Encoders

AIG-CSAM - AI-Generated Child Sexual Abuse Material

AIG-NCII - AI-Generated Non-Consensual Imagery

VLOPs - Very Large Online Platforms

VLOSEs - Very Large Online Search Engines

CDA - Critical Discourse Analysis

Table of Figures:

Preface

My academic journey thus far has been shaped by an interest in the societal implications of emerging technologies and their potential for both advancement and detriment. It was within this broader context that a specific and disturbing application - deepfake sexual violence - came into focus. My focused engagement with deepfake sexual violence began in June 2024, following the widespread online circulation of non-consensual deepfake content featuring Megan Thee Stallion across numerous social media platforms. Although I had prior awareness and pessimism of general AI advancements, deepfake technology presented a unique concern due to its capacity for generating fabricated *intimate* media - an element about which I felt a strong premonition regarding its harms to women. The public discourse surrounding these incidents, where I felt male voices often lacked empathy or recognition, prompted an immediate, in-depth investigation into the technical and societal implications of this rising phenomenon. My discussions with others during this period highlighted a growing concern that this form of digital abuse could soon extend beyond public figures - a foreshadowing that publicly materialised less than two months later.

This premonition became reality with the deepfake crisis in South Korea, which gained momentum by May 2024 and had expanded across the country by August 2024. This sequence of events, though distressing, reinforced the necessity of addressing this specific area of digital harm. Throughout the research period for this thesis, the consistent emergence of large-scale deepfake operations globally - many of which are detailed below - underscored the ongoing and systemic nature of this threat, continuing to solidify my drive to address these digital injustices.

The research process presented significant personal challenges, but in ways I had not anticipated. Investigating these attacks, particularly those involving minors and Child Sexual Abuse Material, exposed me to content and online environments that were unsettling in their scope and nature. Beyond the documented global incidents and continuous reports of new deepfake attacks on women, I directly observed online communities to assess the current landscape of deepfakes, including specific subreddits and Discord servers dedicated to the creation and dissemination of deepfake material. While much of this content depicted animated or fictional characters, these platforms also contained hundreds of deepfakes featuring identifiable celebrities alongside numerous images of seemingly non-public women. Despite the lack of explicit identifying details or definitive proof of their origin, the uncanny verisimilitude in the facial features, especially the eyes, sparked a personal disquiet. This observation, indicating the possible use of real individuals' likenesses, instilled a personal conviction about the urgent need to address consent and privacy in deepfake content, reinforcing the trajectory of my research.

The emotional impact of this exposure, coupled with observations of public and online attitudes that normalised or defended such content, proved a significant aspect of the research experience. My primary challenge shifted from the academic rigor of research, writing, and incorporating feedback, to maintaining a sense of optimism regarding the feasibility of effective

governmental intervention at this stage. The current sociopolitical climate, underscored by a regression in global women's rights, contributes to a personal scepticism that women's digital safety will not be prioritised over economic or technological innovation, especially amidst the competitiveness of global AI. This thesis, therefore, represents not only an academic contribution, but also a dedicated effort to illuminate this critical, unaddressed global issue.

Compiling the following list, documenting major deepfake operations uncovered during my thesis research, was a consistent reminder of the urgency and global scale of this issue. Each new incident reinforced that the South Korean epidemic marked a critical, rather than unique, inflection point, emphasising the imperative to understand and address this evolving threat:

- **Operation Cumberland (Europol-led, February 2025)**: Large-scale operation, spearheaded by Denmark, broke up a global crime group that ran an online platform for the large-scale distribution of AIG-CSAM, leading to 25 arrests across 18 countries: United Kingdom, Germany, France, Italy, The Netherlands, Australia, Austria, Belgium, Bosnia & Herzegovina, Czech Republic, Finland, Hungary, Iceland, New Zealand, Norway, Poland, Sweden, and Switzerland. Resource
- **Spain-led Interpol Operation (June 2025)**: Disrupted numerous large instant messaging groups found circulating vast amounts of CSAM, including AIG-CSAM, resulting in arrests and identified suspects across 12 countries: Argentina, Bolivia, Brazil, Bulgaria, Costa Rica, El Salvador, Honduras, Italy, Panama, Paraguay, Portugal, and the United States. Resource
- **South Korea**: Since late 2024, authorities have continued to uncover and prosecute large-scale deepfake pornography rings, with some operators responsible for creating and distributing thousands of deepfake videos across extensive Telegram chat networks. Resource
- **Canada**: In May 2025, a worldwide investigation revealed a Canadian pharmacist was a key figure behind MrDeepFakes.com, a platform identified as one of the world's most notorious deepfake porn sites, responsible for spreading tens of thousands of non-consensual deepfakes globally before it shut down. Resource
- **United States**: The USA is currently ranked number 1 globally for accessing websites that create sexually explicit deepfakes, with approximately 59.73 million visits. Resource
- **India**: Data from late 2024 showed India was second globally with 24.57 million visits to websites that make explicit deepfakes. Resource
- **Malaysia**: In April 2025, a teenager in Johor was arrested for allegedly creating and selling pornographic AI-made images of dozens of his schoolmates and alumni, distributing them via multiple social media groups. Resource
- **Indonesia**: In May 2025, a university in Bali expelled a student after a deepfake sexual harassment case came to light, where illegal deepfake content of a student was created and widely circulated. Resource

- **Nigeria**: Crime groups, like the ‘Yahoo Boys’, have increasingly used advanced deepfake technology to make convincing fake content for exploitation, with a report in April 2025 detailing how deepfakes are used in large-scale sextortion schemes. Resource
- **Vietnam**: Reports from early 2025 uncovered criminal organisations leveraging deepfake software to superimpose victims' faces onto explicit videos, demanding payment to prevent public release. Resource
- **Japan**: Data from late 2024 indicates Japan ranks third globally in traffic to websites that disseminate sexually explicit deepfake images, with over 18 million visits. Resource

Table of Contents

1. Introduction

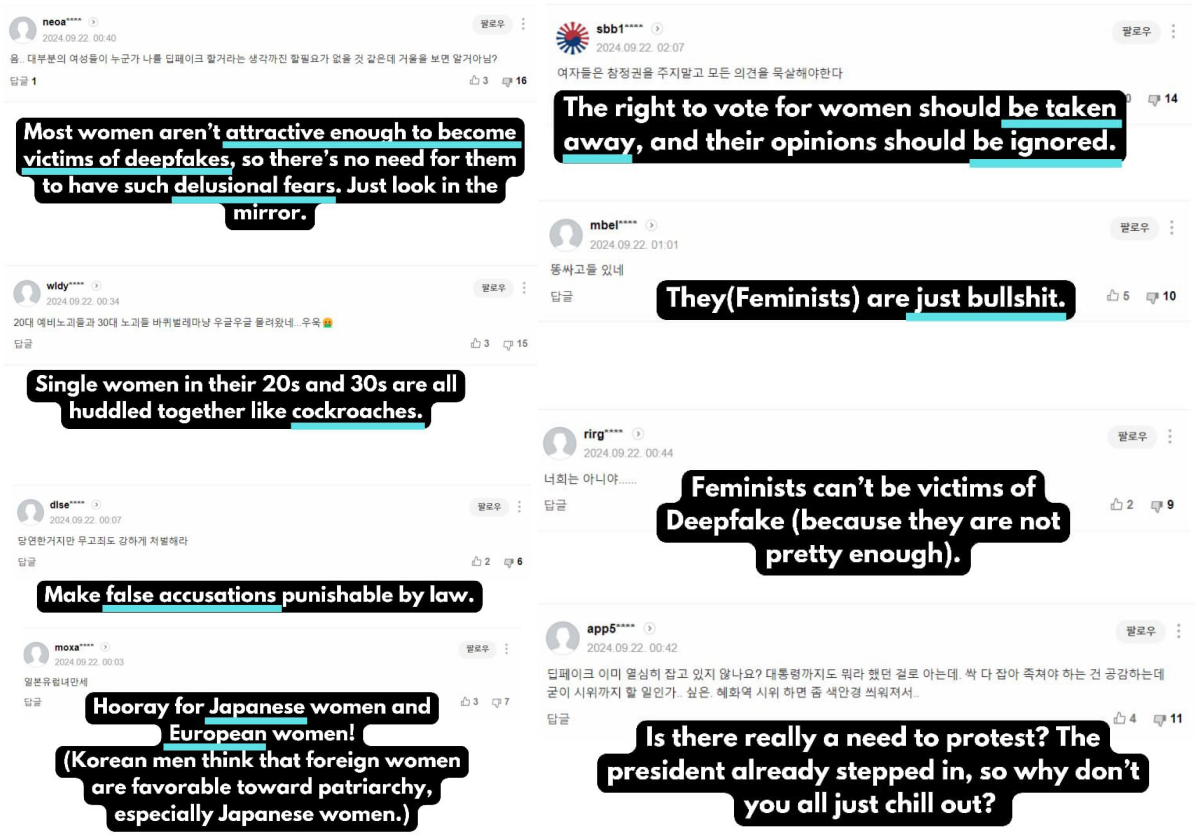

Figure 1: Translated screenshots from a South Korean article (from Reddit, u/SouthKorea_Team_CAT, 2024)

When confronted with such comments [Fig. 1], one's initial assumption might place their origin within an incel forum or a similar extremist platform. Yet, *these* particular sentiments originate from the comment section of a 2024 South Korean article reporting on the large protests of women in South Korea against deepfake abuse. Their presence thus offers a brief manifesto of misogynistic public opinion - an insight into how artificial intelligence (AI) technologies are weaponised against women and young girls, revealing the ingrained societal biases that shape perceptions of these digital harms.

The biases and technological weaponisation evident in these comments stem from the rapid growth of AI. Since its global integration AI's societal impacts have steadily mounted, reshaping creative fields, and driving layoffs across business landscapes (Amidi, 2024, para 2&6; Barrabi, 2025, para.1, 9-10). Despite limited public understanding regarding its workings, society has continued to embrace and integrate AI into many facets of contemporary life (Gordon, 2023, para. 1; Heikkilä, 2024, para.1-4). Unsurprisingly, a new issue has surfaced: AI-generated deepfakes, arriving with a magnitude for which societies and existing institutions appear unprepared (European Commission, 2024, para. 18; Trombevski, 2025, para. 33). This reality was underscored by the

deepfake epidemic that came to light throughout South Korea in 2024 - a revelation that, in hindsight, was somewhat predictable.

This national affront involved the systematic creation and distribution of sexually explicit deepfake images and videos of women and girls without their consent. With an estimated over 10,000 victims, and operations spanning over 400 schools nationwide [Fig.2], the case demonstrated an unprecedented scale and scope, impacting diverse victims and institutions across the nation. For a detailed history, modus operandi, and full scale, see *Appendix A*.

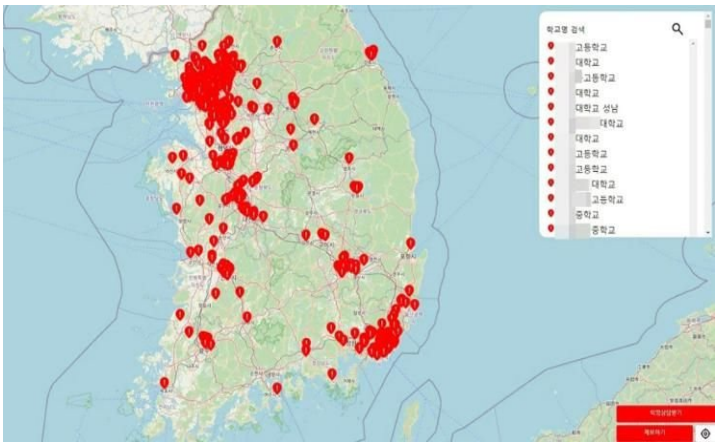

Figure 2: Captured image of a "deepfake distribution map" showing hundreds of schools in Korea that were allegedly affected by the 2024 deepfake crisis (H. Lee, 2024b)

Beyond its quantifiable scale, the South Korean deepfake epidemic revealed a qualitative layer of exploitation and gendered dynamics; while women anxiously deleted their digital presence, some men displayed a lack of remorse or empathy, highlighting a pronounced gendered divide in the perception of this new technology and its harms (W. Lee et al., 2024, para. 5; I. Im, 2024, para 1-10). For specific examples of these responses, see *Appendix A*. The South Korean case offers a solid starting point for discerning broader patterns of deepfake harms. Its manifestations - perpetrator impunity, systemic legal failures, and entrenched misogynistic attitudes - exemplify the architecture of exploitation present in digital spaces, challenging consent dynamics. It is these pervasive phenomena, through the lens of virtual feminist theory, that this thesis will explore in depth to understand their global implications.

Thus, while serving as a cautionary tale, South Korea should not be considered not an anomaly, but a precursor to a broader global issue (European Commission, 2024, para. 18). Deepfake sexual abuse has developed into a pervasive trend, evidenced globally, from viral non-consensual deepfakes of public figures like Taylor Swift in the USA, to a 400% increase in Non-Consensual Intimate Imagery [NCII] abuse in the UK since 2017 (Saner, 2024, para.1; Demony, 2025, para. 4). This rising global trend is compounded by a fragmented international response to AI regulation; major AI actors currently grapple with divergent strategies, signifying a lack of unified governance (Akselrod & Venzke, 2025, para.1; Hendrix & Lima-Strong, 2025, para.1-2; Hemrajani, 2023, para, 1, 10-13). Amidst this, a consistent pattern emerges: a significant global underrepresentation of

female voices in AI development and regulation (Ramos, 2022, para. 9; S. Pal et al., 2024, para. 2). Given the disproportionate amount of harm inflicted almost exclusively upon women, this thesis posits that these two trends are linked, with the biases from a homogeneous masculine creator base embedded in these technologies, enabling their continued, unchecked proliferation and the resultant gendered harm (Jacobsen & Simpson, 2024, p.91-93; Aavik et al., 2024, p.2). This leads to the thesis's primary research question (RQ1): *How are the impacts and perceptions of deepfake technologies related to the underrepresentation of women in AI development and regulation?*

It is this complex landscape of gendered digital violence, differing perceptions of its impacts and harms, and the paradoxical involvement of youth that this thesis seeks to unravel. While the growing body of scholarship addresses deepfake technology, existing research often emphasises algorithmic detection, legal and policy frameworks, or broader ethical discussions. This conventional focus, however, largely overlooks the nuanced gendered dimensions, specific victim experiences, the unique nature of deepfake harms, and paradoxical roles of minors as both victims and perpetrators. While technical studies excel at identifying deepfakes, they do not illuminate *why* certain populations are targeted or *how* those targets experience the resulting digital violence. Highlighted by recent cases like the 2024 South Korean deepfake epidemic, the sensitive dynamics of minor involvement remains significantly underexplored from a critical victim-centred, theoretical standpoint. This thesis argues a qualitative understanding of the specific harms of deepfakes, grounded in virtual feminist and digitalisation theories, is vital for developing effective, human-centred responses that transcend technical or legal solutions. By leveraging a combination of virtual feminist and digitalisation theories, this research addresses this gap, seeking to understand why and how these synthetic media tools disproportionately impact women. Specifically, it examines how societal perceptions of deepfake dangers vary across genders, and how understandings of minors' involvement align with technological accessibility. These insights will be used to philosophically interrogate the inadequacy of existing institutional responses to the sui generis harms of deepfakes, advocating for a more victim-centred governance approach.

The following chapters will explore these issues comprehensively. Chapter 2 establishes the theoretical framework, drawing on virtual feminist and digitalisation theories (Haraway, Butler, Plant, Prensky) to explore the architecture of exploitation and consent dynamics in digital spaces, introducing the key research questions underpinning the research. Following this, the thesis outlines the multi-method qualitative methodology, including focus group research and critical discourse analysis, employed to investigate these phenomena. Chapters 4 and 5 present the empirical results from these investigations, revealing key themes in deepfake perceptions, harms, prevention, and minors' paradoxical role. The subsequent discussion chapter will then interpret these findings through the lens of the theoretical framework, addressing the main research question directly. Finally, the conclusion will summarise the key insights and suggest avenues for future research, contributing to a more equitable and safer digital landscape by illustrating how AI-generated

deepfakes challenge notions of consent, identity, and bodily autonomy for women, and exploring the multifaceted roles of minors within this emerging threat.

2. Theoretical Framework

2.1 The Architecture of Exploitation: Technology, Gender, and Power

While the digital age promised liberation and connectivity, for women it often feels like a new frontier for an old war (van de Hoven et al., 2024, p.4). The history of technological development has been consistently characterised by the systematic marginalisation and exploitation of women, with digital technologies emerging as a contemporary arena for perpetuating long-standing patterns of gender-based control (Plant, 1997; p.10). Challenging technological objectivity, Sadie Plant's (1997) "Zeros and Ones" critically reframes technology as a fundamental mechanisms of social reconfiguration, particularly concerning gender dynamics. Historically, women have been heavily involved with technological processes – from weaving to early forms of programming and electronic assembly – yet their roles were often deemed secondary or supportive, centring on detail-oriented or repetitive microprocesses (Plant, 1997, p.9-12, 74). This positions women differently than men in the space of technological development, who have historically been presented as the authors, inventors, and controllers. Consequently, women's experience with technology, including its potential for harm, cannot be understood simply a mirror image of men's (Plant, 1997, p.38-41).

This historical subjugation of women's technological contributions sets the stage for a critical examination of technological shifts, particularly those of the late 20th century. The 1990s marked a pivotal moment, where a "genderquake", driven by the rise of digital networks, began to reconfigure gender relations, not merely as an extension of existing patterns but as a radical challenge to them (Plant, 1997, p.14-15). This genderquake involved wide-ranging and subtle shifts in sexual differences, relations, identities, and roles, connected to economic changes, where the value placed on physical strength diminished, replaced by demands for speed, intelligence, transferable skills, and communication abilities (Plant, 1997, p.14-15).

As expected, this change was met with resistance; many men expressed emasculation, lamenting that "women and robots had apparently conspired to take their masculinity away" prompting efforts to protect traditional power structures. (Plant, 1997, p.20). This male resistance stemmed from a feeling of losing control as digital technologies facilitated shifts in gender dynamics. Amidst this anxiety, Plant identifies a key manifestation of this resistance: the historical creation of "simulations of the feminine, digital dreamgirls who cannot answer back, pixelated puppets with no strings attached, fantasy figures who do as they are told." (Plant, 1997, p. 80). This drive, she argues, embodies a desire for "Absolute control at the flick of a switch." (Plant, 1997, p.80). Consequently, even as the digital space promised fluidity of identity and potential to bypass traditional constraints like sex or appearance, it also became a site where the anxieties, provoked by the disruption of established gender roles, found new forms of expression and regulation (Plant, 1997, p.25).

Unfortunately, the digital age thus far has not liberated women from their historical oppression but has, instead, transformed the mechanisms of control; technological platforms have become

instruments for reproducing systemic inequalities, offering new architectures of exploitation (J. Kim, 2024, p.3; Haraway, 2016, p.32). From revenge porn and molka, to incel forums and online harassment, women constantly confront increasingly complex forms of technological violence in online spaces; deepfake technologies represent another evolution in digital exploitation, leveraging AI to transform platforms into mechanisms of gender-based violence. (Aavik et al., 2024, p. 2; Zowghi & Bano, 2024, p. 1; Flynn et al., 2021, p.584). Unlike previous forms of digital harassment, deepfakes offer perpetrators an unprecedented tool for psychological and sexual assault, capable of generating hyper-realistic videos that override personal autonomy and dignity with a few algorithmic manipulations – with few punitive repercussions thus far (Aavik et al., 2024, p.1-2; H. Lee, 2024, para 17 & 21).

The South Korean deepfake scandal epitomises this technological weaponisation of gender, revealing how mass digital violence emulates patriarchal social structures and systemic gender tensions (Capello et al., 2023, p.1-3; J. Kim, 2024, p.25, 27). Far from coincidence, the emergence of deepfake sexual-crimes in South Korea, enacted through algorithmic expressions of male control, demonstrate how the country's socio-political environment created a milieu advantageous for digital sexual violence, transforming tools into instruments of gender-based oppression. (J. Kim, 2024, p.25; K. A. Park, 1993, p.140; Chung-Hee, 1993, 85-87). Research into South Korean gender dynamics reveals a society where institutional mechanisms consistently reinforce gender inequality - from educational curricula perpetuating stereotyped gender roles to legal systems marginalising women, South Korea's social infrastructure has long maintained male dominance, evidenced by their rank as 94/146 in the 2024 Global Gender Gap Index (K. A. Park, 1993, p.135&140; J. Kim, 2024, p.25; K. K. Pal et al., 2024, p.325-326)

Against this backdrop of gender inequality and technological weaponisation, understanding the dynamics of escalating social tensions is key. The rise of incel culture is a growing global concern; its manifestation in South Korea offering a salient case study. Here, incel ideology, aligned with patriarchal dynamics, appears more socially accepted, public, and vocal among young male demographics who often perceive societal hostility to their interests. (H. W. Jung, 2023, p.1,2,5,12; Capello et al., 2023, p.1-3). This assertive expression of incel ideology has been met with robust resistance from radical feminist movements; in South Korea, the 4B movement emerged as a radical feminist response challenging traditional gender expectations, intensifying gender tensions and polarising the social landscape, transforming technological platforms became battlegrounds for gender conflict (Makken., 2024; para 1-5; J. Kim, 2024, p.26-27). These two movements, arguably opposite sides of the same coin, reflect a profound gender polarisation; in this environment, men, facing challenges to their masculinity, found in deepfake technologies a means of reasserting control and dominance through algorithmic sexual punishment, actualising Plant's (1997) "digital dreamgirls who cannot answer back" - pixelated puppets embodying absolute male control (Jacobsen & Simpson, 2024, p. 1100-1101; Plant, 1997, p.33, 80). This interaction, underscored in the South

Korean case, is gaining global traction; as the 4B movement and incel culture escalate globally and intensify gender tensions, this aggressive gendered polarisation must be recognised as a foundational element driving digital harms, lest a global repeat occurs (E. T. Kim, 2024, para. 3-6; Shi et al., 2024, p. 3).

In this context, deepfakes occupy a "contested space between rupture and continuity in female objectification," transforming historical mechanisms of control into digital forms of violence (Jacobsen & Simpson, 2024, p.1096-1097). Deepfakes transform [women's] bodies into raw computational material, stripping away agency and reducing individual identities to algorithmically generated fantasies (Jacobsen and Simpson, 2024, p.79-83). As instruments of technological sexual violence, deepfakes inflict harms beyond immediate digital violation, fundamentally challenging personal autonomy and dignity (Flynn et al., 2021, p.15; Jacobsen & Simpson, 2024, p.95). These impacts include psychological trauma (e.g., increased bullying, documented suicides) and widespread erosion of trust, particularly among young women who now view digital spaces as unsafe. (Flynn et al., 2021, p.13; Savino, 2024, para. 3-4; Custers & Fosch-Villaronge, 2022, p.65).

Governmental response further exposed the systemic nature of the problem. For details on South Korea's initial institutional reactions, see *Appendix A*. Aavik et al., argue that AI technologies are not value-neutral, but rather "reflect the values of their creators," often perpetuating systemic inequalities and patriarchal mechanisms (Aavik et al., 2024, p.2). This critical insight - that technology is imbued with existing societal biases - provides another foundation for understanding deepfake technology harms, aligning with Haraway's cyborg theory. Haraway further challenges the view of technological systems as neutral tools, positing them as deeply embedded networks of power, negotiation, and transformative potential (Haraway, 2016, p.27-31, 104). Moving beyond technological determinism, this perspective understands how digital systems, like deepfakes, are fundamentally relational, reflecting and reconstructing societal boundaries, particularly concerning gender and control, encouraging examination of not just *what* deepfakes do, but *how* their development and deployment reflects and reinforces existing power structures (Haraway, 2016, p.75-78). Indeed, the very absence of a robust governmental or judicial responses thus far can be understood as a manifestation of these embedded power dynamics. The state, as a key actor shaping social reality, implicitly or explicitly allows existing patriarchal values to permeate technological spaces, thereby contributing to the perpetuation of digital harm against women (Haraway, 2016, p.166-172). This inaction is not a separate failure, but rather a crucial component of the network of power that technology both shapes and is shaped by, and ought to be scrutinised to the extent that politics and technology are inextricably intertwined in the landscape of digital sexual violence, revealing opportunities for contestation and redefinition within said entangled systems (Haraway, 2016, p. 25, 71-72)

Deepfake sexual crimes are not a phenomenon unique to South Korea, nor the only nation needing scrutiny for its lack of protective AI regulations. Instead, they foreshadow the potential for technological tools to become instruments of systemic oppression when developed without diverse

perspectives and ethical considerations of vulnerable groups online – primarily, women and minors (Custers & Fosch-Villaronge, 2022, p. 109). A 2020 study into NCII [non-consensual intimate imagery] provides context, highlighting the global nature of deepfake pornographic violence: 96% of deepfake videos depicted non-consensual pornographic scenes, with ~99% portraying female victims, and nearly a third of deepfake pornography websites featuring non-Western subjects (Jacobsen & Simpson, 2024, p.91-93). Another 2020 study reported over 100,000 instances of AIG-NCII women created without their knowledge, with some material depicting CSAM [Child Sexual Abuse Material] (Brooks et al., 2021, p.7). Many deepfake creators use an ecosystem of bots on platforms, like Telegram, to facilitate sharing, trading, and selling services associated with deepfake content within the deepfake porn market economy (Koltai et al., 2024, para 12).

In fact, until May 2025, MrDeepFakes, a global AIG-pornography site, operated as a prominent deepfake pornography marketplace, hosting over 43,000 sexual deepfake videos, depicting some 3.8K individuals, collectively watched over 1.5B times (Wise, 2025, para 1-4; Han et al., 2024, p.1). While 95.3% of the targets of these videos were women in the public eye, research showed that, despite having purportedly banned such content, hundreds of videos depicted private individuals as targets, and over 1,000 videos contained violent scenes depicting rape and abuse (Han et al.,2024, p.1). A 2024 study found that the primary motivations behind these deepfake videos included sexual gratification, degradation, and humiliation of their targets. (Umbach et al., 2024, p.12&14)

MrDeepFakes ceased operations in May of 2025, following increased pressure from the United Kingdom and Netherlands who criticised the content as a form of revenge porn (Gulsen & van der Plas, 2025, para.1). However, much like the fabled hydra, when one head is cut off, two take its place. *Candy.ai*, a similar deepfake-based website, allows users to create their own 'AI-girlfriend' (Higgins et al., 2025, para. 24). Their discord server reveals over 2,100 members, with channels dedicated to NSWF content and tips for more realistic sexual deepfakes using the websites 'generate image' function [FIG. 3, 4, 5]. Beyond this, numerous easily-accessible non-consensual deepfake pornographic sites and apps (*Clothoff, Nudify, Undress*, and *DrawNudes)* persist online, manipulating financial and online service providers by disguising their activities to evade crackdowns, continuing to amass tens of millions of users (Koltai et al., 2024, para 4, 5, 12)

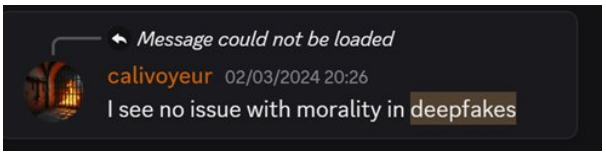

Figure 3: Screenshot from Candy.ai NFSW Discord Server

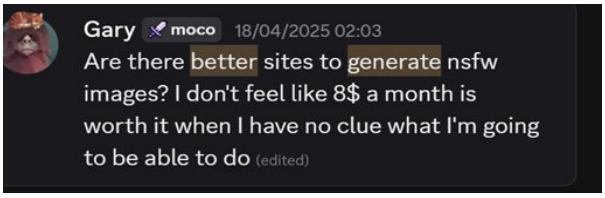

Figure 4: Screenshot from Candy.ai NFSW Discord Server

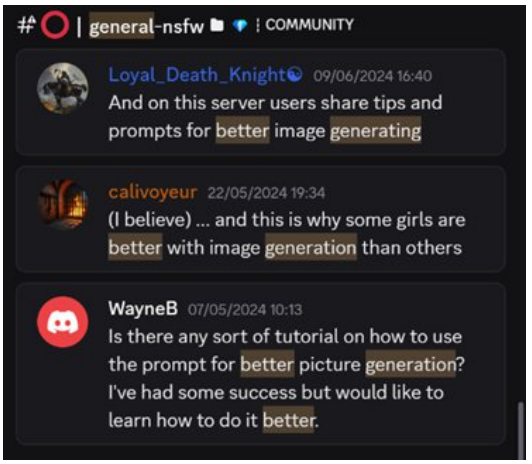

Figure 5: Screenshot from Candy.ai NFSW Discord Server

The unyielding presence of operating openly deepfake platforms, attracting millions of views despite efforts to curb them, underscores a significant and systemic regulatory gap (Koltai et al., 2024, para; 12; Flynn et al., 2021, p.596). This gap, however, extends beyond mere enforcement of existing laws to the technological development, compounded by AI's lack of diversity (Custer & Fosch-Villagrone, 2022, p. 92, 239-241). Zowghi and Bano's (2024) research highlights that AI models predominantly depend on evidence and insight from high-income countries and male academics, with minimal input from minority groups, especially women (p.3-4). This technological monoculture directly contributes to the perpetuation of harmful practices as the inherent biases of a homogeneous creator base are embedded into the very fabric of these systems, allowing for their continued, unchecked proliferation (Custer & Fosch-Villaronge, 2022, p.92, 239, 288).

These technological transformations are, therefore, not accidents but systematic reproductions of entrenched patriarchal logics (Custer & Fosch-Villaronge, 202, p.72, 80-81; Aavik et al., 2024, p.3). They represent a continuation of historical mechanisms of control, transforming patriarchal violence into digitally mediated exploitation that strips women of bodily autonomy and dignity (Flynn et al., 2021, p.15; Jacobsen & Simpson, 2024, p.95). The digital realm has thus become a new frontier where women's autonomy, consent, and bodily integrity are continuously challenged and undermined, revealing technological 'progress' is neither neutral nor emancipatory. (Haraway, 2016, p.27-31; Plant, 1997, p.25). This reality underscores an important oversight: the structures governing AI's creation and control reflect a serious demographic imbalance with studies consistently showing that women comprise only a small fraction of the AI workforce (around 22% globally), with less than 14% in leadership or technical roles (Ramos, 2022, para. 9; S. Pal et al., 2024, para. 2). This distinct lack of female perspectives in AI's development and regulation raises the central question that this thesis aims to answer (RQ1): *How are the impacts and perceptions of deepfake technologies related to the underrepresentation of women in AI development and regulation?*

2.2 Deepfakes as Instruments of Control: Virtual Feminism, and the Dynamics of Consent

Understanding deepfake harms necessitates deconstructing their underlying mechanisms to understand *how* and *why* these synthetic media tools disproportionately impact women. The previous section indicates that the underrepresentation of female voices in AI development has allowed harm to become not just a side effect, but a systemic consequence, failing to address fundamental patriarchal designs perpetuating women's marginalisation. (Custer & Fosch-Villaronge, 202, p.72, 80-81; Aavik et al., 2024, p.3). Deepfake technologies are more than technological innovations; they are complex mechanisms of digital violence exploiting existing power structures and gender inequalities (Flynn et al., 2021, p.15; Jacobsen & Simpson, 2024, p.95). Examining the technical processes enabling hyper-realistic synthetic media creation allows for a more nuanced understanding of the ways consent is violated.

2.2.1 Deepfakes as Mechanisms of Consent Violation

The exponential growth of AI, initially heralded as a transformative breakthrough promising advancements across many domains, has seen rapid acceptance outpace societal understanding and ethical frameworks, creating an uncertain landscape where general digital literacy struggles to keep pace (Grewal et al., 2024, p.939 & 945; Brooks et al., 2021, p.32-33). AI is often misconceived as a monolithic entity; rather, it is an umbrella term for a broad ecosystem of technologies encompassing machine learning, neural networks, and synthetic media production (Brooks et al., 2021, p.4; Brown, 2021, para. 3-4). Within this, deepfakes represent a controversial manifestation of this technological convergence: a form of synthetic media that uses AI to superimpose or manipulate human likenesses with unprecedented realism (Vecchietti et al., 2024, p.2; Buffet Brief, 2023, p.1).

Combining “deep learning” and “fake media”, deepfake technologies represent a combination of AI, machine learning, and digital manipulation, radically challenges existing paradigms of identity, consent, and bodily autonomy (namuwiki, 2025, para.1; Mahmud & Sharmin, 2020, p.13; Brooks et al., 2021, p.3; Vecchietti et al., 2024, p.3). Deepfakes are thus defined as synthesised media that digitally alters an individual's likeness through deep generative methods, utilising techniques from AI to manipulate visual and audio content with realism (Mahmud & Sharmin, 2020, p.13). For a detailed technical explanation of deepfake architectures and generation processes see *Appendix B*.

Crucially, the foundational data for deepfake creation, often requiring extensive video samples to capture comprehensive facial characteristics, is frequently acquired without explicit consent or knowledge of the individual depicted (Brooks et al., 2021, p.17; Kobis et al., 2024, p.2). This non-consensual data acquisition lays the groundwork for subsequent non-consensual manipulation, also understood as Non-Consensual Intimate Images (NCII).

The rapid evolution of deepfakes has significantly eroded the reliability of human perception (Charlwood, 2025, para 2). Unlike traditional image manipulation, AI systems can now overcome prior limitations; where earlier iterations initially struggled with complex features like eyes, ears, and hands, current versions of AI deepfake technology can generate uncanny representations by understanding subtle contextual cues and facial nuances, 'learning' from mistakes (Mahmud & Sharmin, 2020, p.16-18). This self-improving design, coupled with increasing accessibility and democratisation of deepfake creation tools, poses formidable challenges for distinguishing synthetic from authentic media as the technology matures. See *Appendix B* for further technical details on these advancements and specific software.

2.2.2 Deepfake technology at the Intersection of Virtual Feminism

The implications of this democratised, rapidly advancing technology extends beyond digital manipulation, revealing an important mechanism of identity reconstruction (Aavik et al., 2024, p.1-2; Haraway, 2016, p.33). These synthetic representations function as digital doppelgangers, exposing the performative nature of identity - particularly gender - by revealing the constructed and malleable boundaries of bodily representation (Plant, 1997, p.177; Buffet Brief, 2023, p.1). Here, the digital body becomes a site of radical recontextualization as deepfakes do not simply reproduce an image but actively reconstruct identity via algorithmic intervention (Buffet Brief, 2023, p.1; Butler, 1999, p.174-175). Like the performative nature of drag, these synthetic representations reveal the fundamental instability of gender as a fixed category (Butler, 1999, p.174-175). The technology exposes how identity is not an immutable essence, but a series of repeated acts, digitally manipulable, reconfigurable, and weaponizable.

This capacity for weaponised identity reconstruction illuminates the severity of harm deepfakes can inflict upon women. At its core, this harm stems from their violent intervention in the performative construction of gender, leveraging digital network fluidity to enforce patriarchal control (Aavik et al., 2024, p.2). This multiplex reality necessitates a theoretical approach accounting for the intricate ways deepfakes operate at the intersection of technology, gender, and power. To that end, the combined insights of Donna Haraway's (2016) cyborg framework, Judith Butler's (1999) theory of gender performativity, and Sadie Plant's (1997) analysis of digital networks are leveraged to dissect the gendered dimensions of deepfake harms and understand why women experience this technology so distinctly.

Donna Haraway's (2016) cyborg, an ironic political myth blending organism and machine, blurs previously distinct boundaries and embodies the inextricable intertwining of bodies and technologies in modern life; deepfakes embody this confusion perfectly, blurring the lines between authentic persons, their digital representations, and fabricated actions (p. 5-7, 33, 43). Haraway's contention that technologies actively embody and enforce new social relations, particularly for women, contextualises deepfakes as instruments of social reconfiguration within what she terms the "informatics of domination." (Haraway, 2016, p.33). This concept describes a new matrix of power

and social control operating through communication technologies, information systems, and biotechnologies, replacing older forms of industrial control with power exerted through data, codes, and networks (Haraway, 2016, p.33 & 46). Deepfakes exemplify this, leveraging algorithmic control over digital information (images, videos) to exert power - often in gendered and oppressive ways - making their creation and dissemination a direct act within this system of informatics. (Aavik et al., 2024, p.2; Haraway, 2016, p.29-30). Furthermore, while Haraway's cyborg might experience "pleasure in the confusion of boundaries" - a liberatory potential derived from the blurring of human/machine, organism/text, or male/female distinctions, offering escape from rigid categories and dominant narratives towards a more fluid and agentic identity – deepfakes immediately pervert this utopian vision (Haraway, 2016, p.1). Rather than fostering agency and transformation through hybridity, the seamless blurring of real and fabricated boundaries by deepfakes becomes a tool for harm: the digital self is hybridised, not by choice for emancipation, but by force for violation, often for sexual gratification (Haraway, 2016, p.65; Umbach et al., 2024, p. 12). The very fluidity that could be liberating is exploited, becoming a source of vulnerability and distress for those - predominantly women - whose digital and personal boundaries are violated.

Building on Haraway's (2016) framework of technology as a site of power, Judith Butler's (1999) theory of gender performativity illuminates *how* deepfakes inflict harm within this technosocial matrix. Her theory posits that gender is not an internal essence but is manufactured through a "sustained set of acts, a repetition and a ritual" (Butler, 1990, p.xv). What appears as an interior essence of gender is manufactured, and the appearance of substance, of a natural sort of being, is an effect of this performance. Deepfakes hijack performed gender construction: imposing simulated, fabricated acts onto a [female's] digital likeness (Brooks et al., 2021, p.1; Aavnik et all., 2024, p.1-2). This is not mere misrepresentation, but a violent imposition of a false performative act, creating a digital stylisation of the body that is alien to the individual's lived reality (Butler, 1999, p.29-31). By fabricating a performance, deepfakes directly attack the lived social and bodily realities of women and girls, who comprise 96% of AIG-NCII victims, turning their gendered performance - a vital strategy of survival within compulsory systems - against them (Jacobsen & Simpson, 2024, p.91-93; Butler, 1999, p.177-178). Butler also highlights that the body's boundaries are politically constructed and maintained; deepfakes further exploit this wherein the simulated performance breaches the digitally represented bodily boundary, leveraging existing systems of meaning and power that have historically inscribed and controlled women's bodies (Butler, 1990, p. 26-27; Aavnik et al., 2024, p.2).

This performative violence, enacted upon the digitally constructed self, is amplified and disseminated within the fluid digital space Plant describes. Plant highlights the digital matrix, the integrated circuit, and the network as the operative space of contemporary power and identity formation - the very infrastructure where Haraway's informatics of domination shapes how power is exercised and identity is negotiated (Plant, 1997, p.58-60). Plant points to the inherent fluidity and

permeability within the digital space, characterised by "criss-crossing the complex topical landscape" and establishing "multiple connections" - the non-hierarchical, decentralised, and interconnected nature of digital networks, where information flows in myriad directions, linking disparate points rather than following rigid, linear paths (Plant, 1997, p.11). She further associates this network fluidity with historically gendered concepts like "hysterical" or "fluid genetic transfers", suggesting a deep historical resonance between the female experience and the characteristics of digital networks, emphasising that qualities traditionally associated with and often pathologized in women - such as fluidity, non-linearity, and resistance to singular control[1]- are precisely the qualities that define the dynamic, uncontrollable nature of digital networks themselves (Plant, 1995, p.109-111, 170, 241-247). This perspective suggests that the attributes that make networks powerful and disruptive to old, patriarchal orders are the same attributes historically linked to and feared in women. The provocative cyberfeminist manifesto statement - "The clitoris is a direct line to the matrix" - encapsulates this revolutionary vision, suggesting digital spaces are embodied territories where traditional gender binaries could, in theory, be destabilised (Plant, 1997, p.59). Deepfakes, however, weaponise this potential, being that they are not simply products of, but dynamic processes that exist and spread within this fluid digital network. Their ease of creation, copying[2], and dissemination is a direct function of this digital permeability of boundaries and the transformation of commodities (including images) into digital form (Ajder et al., 2019, p.4; Mahmud & Sharmin, 2020, p.19; Plant, 1995, p.14, 35). The digital network facilitates the rapid, widespread dissemination of deepfakes, making violations of performatively constructed identity and politically signified bodies incredibly difficult to contain or control once released into the digital sphere; this uncontained flow, facilitated by the network's fluidity, allows pervasive enforcement of meanings and control of the ideal woman through the creation and spread of these manipulative simulations (Mahmud & Sharmin, 2020, p.19; Plant, 1997, p.143-154).

This fluid, networked reality creates vulnerabilities; women's position within the integrated circuit, intertwined with issues of sexuality and precarity, makes them susceptible to the harms enabled by the fluid, networked nature of deepfakes; a susceptibility compounded for younger individuals lacking comprehensive digital literacy or avenues for recourse (Haraway, 2016, p.32; Aavik et al., 2024, p.2; Ajder et al., 2019, p.2). By synthesising these perspectives, it becomes evident that deepfakes are little more than performative simulations (Butler, 1999) weaponised within the fluid digital networks (Plant, 1997) of Haraway's (2016, p.29-33) informatics of domination. They specifically target [women's] lived realities by attacking the performatively constituted self and its politically constructed bodily boundaries with differential harm stemming

[1] E.g. Historical notions of "hysteria" challenging patriarchal rationality, or the idea of shifting "genetic transfers" destabilising fixed biological identities (Plant, 1995).

[2] As simulacra - referring to a copy or imitation that does not have an original, or that has lost its connection to the original. In other words, it is a copy without a "real" referent (Massumi, 1987)

from how these simulation technologies and digital fluidity exploit and amplify existing gendered norms and power structures onto women as the primary targets (Jacobsen & Simpson, 2024, p.91). The network enables rapid, uncontrollable imposition of fabricated, often sexualised, gendered performance onto a woman's digital image, leveraging her historical vulnerability as a contested site of meaning and control within the integrated circuit. This integrated theoretical framework thus exposes the complex, gendered threat deepfakes pose, fundamentally compromising women's and young girls' autonomy and safety in the digital age.

Given this understanding of deepfake harms and their distinct gendered implications, a critical sub-question arises regarding their broader societal reception. Effectively addressing such pervasive digital threats often hinges on public awareness and how these dangers are understood across demographics. Therefore, to grasp the societal landscape shaped by deepfake proliferation, this thesis poses a second research question (RQ2): *How do perceptions of the dangers of AI-generated deepfakes fundamentally differ between men and women?*

2.3 The Generational Nexus: Minors, Media, and the Breakdown of Binaries

While discussions of deepfake technologies have centred on the victimisation of adult women, evidence reveals this harm is not limited by age (Wolbers et al., 2025, p.1-2; Theil et al., 2023, 7-8). This section explores minors' complex, dual role within the deepfake landscape - as highly vulnerable victims and, paradoxically, a significant demographic of perpetrators. It also examines how media coverage, potentially reflecting a "digital immigrant" gap in understanding the nuances of the technology and its harms, may influence societal perceptions of the severity and magnitude of their involvement.

NCII statistics expose an ecosystem where children are not merely collateral damage, but primary targets of advancing AI-facilitated abuse [AIFA] (Wolbers et al., 2025, p.1-2; Theil et al., 2023, 7-8). The scale of this issue is significant: The Internet Watch Foundation's (2024) report revealed an escalation in AIG-CSAM, documenting over 3,500 new criminal images uploaded to a single dark web forum in one reporting period (Theil et al., 2023, p.7). The report further highlights AI's capability to generate highly realistic videos depicting child sexual abuse with unprecedented realism; around 90% of assessed AIG-CSAM images appear realistic enough to be assessed under 'real' CSAM laws (Internet Watch Foundation, 2024, p.7). Additional statistical evidence, from 1,040 students aged 9-17, revealed that 11% of students knew classmates who had used AI to generate NCII of their peers, with another 10% unwilling to disclose such information, suggesting that between 1 in 9 and 1 in 5 students are aware of AIFA within their immediate social circles (Wolbers et al., 2025, p.2; Healey, 2024, para. 4-7).

Global incidents underscore this widespread crisis, directly challenging the prior, almost exclusive, focus on adult women as victims; an emphasis that has echoed into academic discourse,

exposing a research gap on this dynamic (Atherton, 2024, para 4-10; K. Williams, 2024b, para.1; Koltai & Zhu, 2024, para. 29; Wolbers et al., 2025, p.5-7, 10-11). For instance, South Korea's 2024 deepfake case exposed hundreds of organised networks targeting school children, from elementary to high schools, through mass deepfake creation and dissemination, highlighting unforeseen minor involvement in such crimes (J. Jo et al., 2024, para. 1; Hankyoreh, 2024, para. 2). While an extreme example, similar patterns have echoed worldwide – from New Jersey's Westfield High School, where teen boys created sexually explicit deepfakes of female classmates, to incidents in California and Illinois using AI-powered 'nudification' apps to manipulate images of underage girls (Cruz, 2024, para.37; Singer, 2024b, para. 13; Atherton, 2024, para.10). The technological infrastructure itself has become complicit; research indicates that some AI models are potentially trained on existing child sexual abuse material, creating a self-perpetuating cycle of victimisation (Thorn, 2025b, p. 10)

In fact, foundational datasets, like LAION-5B[3], have been exposed to contain numerous instances of known and new CSAM (Theil, 2023, p.2; Thorn, 2025b, p.6). Perpetrators can then exploit these contaminated models by fine-tuning them using Low-rank Adaptations (LoRAs) to optimise model weights specifically for generating more abusive content, potentially creating a self-perpetuating cycle where existing abuse material is used to generate new synthetic abuse imagery (Theil 2023, p.5-6; Thorn, 2025b, p.3-4). OpenDream, a generic AI image generation site, exemplifies this concern in practice: its introduction of NSFW[4] prompts and models allowed users with minimal technical expertise to generate explicit content (Koltai & Zhu, 2024, para. 13). This directly coincided with a surge in attempts to create AIG-CSAM and NCII, demonstrating how easily accessible AI tools translate model vulnerabilities into widespread, facilitated abuse, often bypassing safeguards and contributing to an ecosystem where this material can even be monetised. Consequently, as these AIG-CSAM become hyper-realistic and virtually indistinguishable from real content, their presence extends beyond specific platforms. Critically, search engines like Google and Bing have inadvertently indexed and returned synthetic sexualised images of minors, including bikini-clad and naked toddlers, making them readily discoverable across the public internet (Koltai & Zhu, 2024, para. 11). What emerges is not an isolated phenomenon, but a systemic technological crisis that transforms children's digital spaces into potential sites of sexual violence (Aavik et al., 2024, p.2). The combined ease of image generation with anonymity provided by digital platforms has created an unprecedented landscape of vulnerability for minors - simultaneously the most

[3] LAION-5B (Large-scale Artificial Intelligence Open Network) is a large, open-source dataset consisting of 5.85 billion CLIP-filtered image-text pairs (Beaumont, 2022). It was created to democratise research and enable the training of large-scale multi-modal artificial intelligence models. Stable Diffusion, a well-known text-to-image model, was trained on a 2 billion-image subset of LAION-5B (Salvaggio, 2024).

[4] An internet acronym for Not Safe For Work, indicating content that is explicit, offensive, or otherwise inappropriate for viewing in a professional or public setting (Cambridge Dictionary, 2025)

technologically fluent yet most unprotected demographic (K. Williams, 2024b, para. 9-12; Theil et al., 2023, p.8-9).

This statistical reality necessitates confronting the traditional victim/perpetrator binary: minors are not only among the most vulnerable victims but also constitute a large demographic of deepfake perpetrators. South Korea's nationwide epidemic, for example not only highlighted the extensive victimisation of minors but also revealed that a substantial portion of the offenders were minors themselves; for detailed statistics on minor involvement, see *Appendix A*. This byzantine dynamic left law enforcement and legal systems struggling to effectively tackle an issue where the lines between the exploited and the exploiter are blurred (Koltai & Zhu, 2024, para. 31-32; Custers & Fosch-Villaronge, 2022, p.216). Even after its revelation, these statistics continue to increase, highlighting a serious inability to tackle and prevent this issue (J.-Y. Choi, 2025, para, 2-5). This global phenomenon, evidenced by widespread incidents in schools and online communities, underscores minors' unique and paradoxical position within the deepfake ecosystem.

To grasp minors' unique position – how it renders them especially vulnerable while also affording a dangerous degree of reckless freedom - Marc Prensky's "digital native" concept offers insights[5]. Prensky (2001) argues that individuals born into a digitally saturated world possess innate fluency with digital tools and networked environments, fundamentally differing them from the "digital immigrants" who adopted these technologies later in life (p.1-2). This inherent comfort with digital interaction and information processing, though advantageous, simultaneously contributes to minors' deepfake-related vulnerability, affording them a dangerous degree of reckless freedom (Prensky, 2001, p.2).

2.3.1 Digital Natives as Victims and as Perpetrators

The qualities defining digital natives can heighten their susceptibility to deepfake victimisation. While adept at navigating online spaces, their digital fluency often lacks comprehensive critical media literacy or a full understanding of online risks (Prensky, 2001, p.2; Krasna & Bratina, 2011, p.1249). This digital generation, accustomed to consuming and creating vast amounts of online content, may lack critical faculties to identify realistic AI-generated fabrications from authentic media (Charlwood, 2025, para 2). Deepfakes, with their uncanny realism, exploit this gap, leaving minors vulnerable to manipulated content. Furthermore, minors are in a critical developmental stage where identity formation is fluid and heavily influenced by social validation and

[5] Prensky's digital native concept, while influential, is not without critique. Contemporary scholarship recognises that technological fluency is more nuanced than a simple generational binary, influenced by factors like socioeconomic access, educational opportunities, and individual technological engagement (Farkas & Maloney, 2024; Garg & Sengupta, 2019). This research uses the concept as an analytical framework while acknowledging its limitations in fully capturing the complex relationship between individuals and evolving digital technologies, especially on a global scale.

peer perceptions; deepfakes, especially if sexualised or humiliating, directly attack this developing sense of self, leading to psychological trauma, reputational harm, and long-term emotional distress (Elkind, 1967, p.1026-1030; Singer, 2024a, para.9). Their extensive online footprints - primarily their social media activity - provide an abundant, often unknowingly contributed, source material for deepfake creators, exacerbating their exposure. The desire for peer acceptance, coupled with still-developing emotional regulation skills, can also leave minors susceptible to deepfake-enabled sextortion or coerced participation in risky online behaviours (Krasna & Bratina, 2011, p.1249).

Conversley, the digital native framework also explains minors' prevalent role as deepfake perpetrators. Their inherent comfort with digital manipulation and seamless integration of technology into daily life can foster an environment where creating deepfakes is perceived as an extension of familiar online activities like photo editing or meme creation (Prensky, 2001, p.1). Indeed, it has been suggested that teenagers involved in the perpetration of these crimes often view it as little more than an old common prank (Y. Choi, 2024d, para. 7). The accessibility of user-friendly deepfake generation tools, requiring minimal technical expertise, lowers the barrier to entry for young individuals adept at quickly mastering new digital interfaces; this ease, however, often outpaces the development of ethical reasoning and a full comprehension of real-world consequences (Healey, 2024, para.3; J.-J. Lee, 2024, para 17&30). The perceived anonymity of the internet, coupled with immediate gratification from online attention or peer validation, can diminish the perceived severity of actions (J.-J. Lee, 2024, para 30-31; K. Williams, 2024b, para. 8-9). Moreover, within encrypted chat groups and online communities, seen within the Telegram networks, a culture of impunity can emerge, normalising the creation and sharing of NCII, even AIG-CSAM, manufacturing an environment where harmful content proliferates outside traditional ethical boundaries or adult oversight (Internet Watch Foundation, 2024, p. 15&17). The digital native's inclination for rapid content sharing and multitasking can inadvertently contribute to the wide dissemination of deepfakes, often without full comprehension of the lasting damage inflicted (Prensky, 2001, p.1-3; Y. Choi, 2024d, para. 7). Given these complex dynamics of accessibility and engagement, a question emerges regarding public understanding of this issue (RQ3): *How do perceptions regarding the use and accessibility of deepfake technologies by and against minors reflect the actual accessibility?*

2.3.2 The Digital Immigrant Impact in Media Representation and Regulation

While digital natives navigate this complex landscape, the institutions tasked with regulating and responding to deepfake crimes - legal, judicial, and regulatory bodies - are predominantly "digital immigrants" (Prensky, 2001, p.2). These individuals, having adopted technology later in life, often retain a "digital accent", manifesting as a fundamental lack of understanding of online behaviours, digital cultures, and networked technologies' fluid nature (Prensky, 2001, p.2). This epistemological gap impacts their ability to formulate effective ethical regulations around deepfake technology and comprehensively address minors' involvement; rooted in analogue paradigms, their

conceptual frameworks struggle to fully grasp the magnitude and severity of harms unfolding in decentralised, rapidly evolving digital spaces, where traditional notions of evidence, jurisdiction, and consequence are challenged by algorithmic speed, global dissemination, and digital content's permenance (Prensky, 2001, p.4&6; Custer & Fosch-Villaronge, 2022, p.213-214, 552, 558; K. Williams, 2024b, para. 12).

The challenge, however, extends beyond a generational understanding gap. Despite considerable data on emerging deepfake harms, public discourse and decisive action remain scant. This raises an important challenge: why does ignorance persists despite readily available information? This thesis posits sustained ignorance is not a passive failing but a systemic outcome, intertwined with larger power dynamics. Digital immigrants, lacking inherent technological fluency, often rely on external sources like news to comprehend complex digital realities and their harms. Yet, when media representation itself carries a "digital accent" - the blind leading the blind - it perpetuates a fundamental misunderstanding of deepfakes' true nature and severity. This media-mediated epistemic gap, arguably, serves broader patriarchal interests, as deepfake crimes disproportionately impact women and girls, implicitly de-prioritising urgent attention. It also aligns with capitalist imperatives prioritising unchecked innovation over robust, potentially restrictive, regulation. This dynamic of obscured understanding ultimately hinders effective policymaking and response, making detailed analysis of the digital accent in news coverage critical to unravelling the lack of current action.

News coverage, often filtered through the lens of digital immigrant perspectives, may struggle to convey these crimes' full depth and nuance, oversimplifying multifaceted online behaviours like peer pressure, online validation, or blurred digital consent to mere malice or naivety (Tenor & Himma-Kadakas, 2023, p.137-139 & 153; Siepp, 2023, p.396). Furthermore, media often focuses on sensational aspects rather than systemic issues, prioritising shock value over a deeper analysis of the underlying technological affordances, platform accountabilities, or deficiencies in digital literacy education (Siepp, 2023, p.396; Atherton, 2024, para.12). This fragmented portrayal contributes to a broader lack of public awareness regarding the true nature of this issue among minors, failing to communicate the severity of the impacts of deepfakes on young victims (Siepp, 2023, p.393). Consequently, this leads to a shortfall in robust protective regulations. By misrepresenting the problem's true scope, media narratives inadvertently leave minor victims insufficiently protected and minor perpetrators inadequately addressed, failing to account for their unique developmental and digital contexts (Tenor & Himma-Kadakas, 2023, p. 146-148, 153, 193; Singer, 2024b, para.12-14; K. Williams, 2024a, para.1). The sudden emergence of deepfakes is catching many institutions unprepared, resulting in precarious safeguards for students; many school districts remain confused how to even report such incidents (Singer, 2024a, para. 11-12). The disconnect between the digital native's lived reality and the digital immigrant's interpretive framework thus profoundly hinders effective societal response and safeguard implementation. This analysis underscores the media's

critical role in shaping public understanding and, subsequently, policy. Given the impact of these representations on how society, legal systems, and educational institutions respond to deepfake crimes involving minors,, a crucial fourth research question for this thesis emerges (RQ4): *How do media representations of AI-generated deepfake crimes relate to minors as both victims and perpetrators?*

3. Methods

Grounded in feminist standpoint epistemology, reflecting its virtual feminist framework, this research adopts a critical theoretical framework, emerging from systemic knowledge production failures in understanding women's lived experiences (Harding, 2016, p.105). Harding (2016, p.105) argues that traditional scientific methodologies inherently privilege perspectives systematically marginalising women's experiences, particularly where technological power intersects with gender dynamics. The choice of standpoint epistemology is thus not merely methodological, but inherently political. Traditional research methodologies consistently reproduce androcentric biases, creating an "excessively weak" conception of objectivity that fails to critique underlying power structures embedded in technological development (p. 111, 116-17, 119-120). Standpoint theory offers an alternative by positioning marginalised experiences as epistemologically privileged, recognising that those experiencing systematic oppression often possess unique insights into social structures precisely because their marginalisation requires a deeper understanding of power dynamics (Harding, 2016, p.119-120). In the context of deepfake technologies, this means centring women's experiences not as peripheral data points, but as key knowledge generators revealing the underlying mechanisms of technological violence (Harding, 2016, p.123-124).

Strong objectivity is crucial; unlike traditional notions claiming neutrality, it acknowledges knowledge is socially situated (Harding, 2016, p.138-140). This demands critical examination of the social and political contexts shaping knowledge production (Harding, 2016, p.149-150). For deepfake research, this means explicitly interrogating how patriarchal technological ecosystems generate and perpetuate harm, rather than treating technologies as neutral tools. This epistemological stance is vital, given its focus on technological violence where women's oppression generates what Harding describes as a “fewer interests in ignorance" perspective – a lens less invested in maintaining existing power structures and norms (Harding, 2016, p.150). For deepfake technologies, this translates to a research approach that does not only describes harm but actively seeks to deconstruct the systemic mechanisms enabling such violence. This approach does not claim universal truth, but instead offers a more nuanced, contextually rich understanding of how technological systems reproduce and amplify gender-based violence.

Engaging with this perspective, the research adopts a multi-method qualitative approach triangulating focus group research (Casey & Krueger, 2015, p. 20; Kitzinger, 1995, p. 3) and critical discourse analysis (Van Dijk, 1998, p. 4; Wodak & Meyer, 2015, p. 18) to thoroughly investigate AI-generated deepfake technologies' gendered dimensions. This strategy explores the intersections of technological harm, gender dynamics, and media portrayals, focusing on the experiences of women and minors. The objective is to explore and understand deepfake violence landscapes by generating data revealing how gender, age, and digital literacy intersect with deepfake harm experiences and responses. Employing both qualitative methods, this research moves beyond surface-level descriptions, producing a transformative understanding of how technological systems reproduce and

amplify gender-based violence. Triangulating these approaches ensures a comprehensive, multi-dimensional exploration of the research questions, providing depth, breadth, and critical analytical rigour to investigating deepfake technologies and their gendered implications.

3.1 Focus Group Design and Implementation

To address RQ2 and RQ3, concerning gendered perceptions of deepfake harms and understanding perceptions of accessibility by and against minors, this research employed qualitative focus groups. Drawing from Morgan's (1996) conception, this method extends beyond individual opinions, instead understanding how meanings are collectively constructed through interactive dialogue (p.130 & 139-140). This group setting allowed participants to challenge, build upon, and negotiate perspectives in real-time, offering a more dynamic data generation process than individual interviews (Kitzinger, 1995, p.299; Morgan, 1996, p.139-140). This approach recognises that group interactions can reveal more than individual perspectives, shedding light on shared understandings, priorities, language, while encouraging participants to generate and explore their own analysis of common experiences (Kitzinger, 1995, p.302). This dynamic process was crucial for understanding how gender shapes deepfake danger perceptions, uncovering differing conceptual frameworks that inform how men and women articulate their views (Casey & Krueger, 2015, p. 38). By encouraging participants to explore their own analyses, the method generated insights into shared understandings, common misconceptions, and the societal frameworks through which deepfake's harms and the roles of minors are perceived.

Participant recruitment targeted young adults with varying familiarity of deepfake technologies. Key criteria included: individuals aged 18+; currently studying or recently graduated (within 18 months); and no specialised technological knowledge of deepfakes. While posing several challenges, recruitment resulted in 14 participants.[6]

The split gendered design intentionally mimicked the gendered dynamics observed in deepfake-proliferating digital spaces, referencing South Korea's Telegram dynamics. This decision was grounded in research demonstrating how gender dynamics significantly influence conversational patterns (Benstien Miller, 2015, p.62). Moreover, previous studies show that gender composition can dramatically affect discussions of sensitive topics, particularly those involving power, technology,

[6] Initial approaches - a recruitment survey via social media platforms (Facebook, WhatsApp, Instagram) and contacting university-affiliated student groups - yielded limited responses, often related to the in-person nature and availability needed for focus group participation. Overcoming this, a refined recruitment strategy was implemented incorporating in-person recruitment, direct faculty email outreach, and snowball sampling from initial participants. This revised strategy recruited 14 participants: 5 via WhatsApp messages, 2 via faculty emails, 2 through in-person recruitment, and 5 through snowball sampling.

and social dynamics (Kitzinger, 1995, p.301; Casey & Krueger, 2015, p.122; Montemurro et al., 2014, p.140; Kollock et al., 1985, p.1). This design allowed exploration of differing discourse patterns within single-gender versus mixed-group settings. Participants indicating male gender preference were offered availability for male-only and mixed-gender groups; those indicating female gender preference were offered availability for female-only and mixed-gender groups; and those selecting non-binary, "prefer not to say," or "third gender" were offered the mixed-gender option only. This stratification, while focused on male/female dynamics to align with the literature's predominant gendered nature of deepfakes, allowed for broader representation within the mixed group.

Data Collection

Three distinct group configurations were conducted: Control Group (Mixed-Gender): 3 males, 4 females (65 minutes); Male-only Group: 4 males (75 minutes); Female-only Group: 3 females (82 minutes).

All groups convened in a neutral setting, curated to encourage genuine dialogue, with chairs arranged in a relaxed circle. To minimise researcher influence and encourage authentic participant interaction, a non-directive approach was adopted, aligning with Casey and Krueger's (2015) recommendations (p. 20, 23). The researcher acted primarily as a facilitator, guiding the discussion while minimising personal input and interpretations, thus upholding principles of grounded theory emphasising emergent data and participant-centred perspectives (Casey & Krueger, 2015, p. 150). Minimal probing was employed, prioritising open-ended questions to encourage spontaneous and in-depth discussions (Kitzinger, 1994, p. 173; Casey & Krueger, 2015, p. 187). A flexible discussion guide was developed, guided by the theoretical framework, incorporated three distinct deepfake scenarios [adult woman, adult man, and a minor] to gauge perceptions of these crimes. Other questions spanned: the perception of harm in different scenarios, including unique victim impacts; contributing social and systemic structures; differences in harm between generation and distribution; the dual role of minors as victims and perpetrators; and institutional or societal responsibility and prevention (*Appendix E*). These prompts encouraged participants to explore their perceptions, fears, and understanding of deepfake harms from individual perspectives.

Rigorous ethical protocols were implemented throughout, including informed consent forms signed by all participants prior to involvement. The study's purpose, participant rights - including the explicit ability to withdraw without penalty - data confidentiality, and the voluntary nature of participation were clearly explained both at the beginning and end of the study. Confidentiality was strictly maintained, protecting participant anonymity through pseudonyms and secure data storage (Krueger & Casey 2015, p. 47,50-54, 62-64). Given the sensitive nature of the subject matter (NCII and AI-G CSAM), participants were heavily prefaced with warnings. It was reiterated throughout that they retained the right to not answer or to withdraw from the study if uncomfortable; no participant ultimately withdrew.

All focus group discussions were audio-recorded with participant consent. Immediately upon completion, participants were debriefed on research's aim and reminded of their ongoing rights. Preliminary notes were made directly after each discussion to capture initial observations and contextual details, compiling the main researcher's notes with the notetaker's, who focused on body language changes. Detailed transcripts were later created from the audio recordings, ensuring accuracy and clarity for subsequent analysis.

Data Analysis

Grounded theory guided the iterative analysis, involving coding and categorising the transcribed data to identify key themes and patterns emerging directly from participant narratives (Glaser & Strauss, 1967, p. 1, 44-45; Krueger & Casey, 2015, p.203). Analysis focused on identifying similarities and differences in perceptions across the three groups (female, male, mixed gender). Focus group analysis commenced with open coding each transcript, identifying initial concepts, ideas, experiences, and actions by fracturing data for conceptualisation. From these initial concepts, categories and their properties were identified using the constant comparative method (Glaser & Strauss, 1967, p.105-115). Attention was paid to how participants perceived the risks and dangers posed by AIG-NCII, whether distinct gendered perspectives emerged around the potential harms of deepfakes, and what solutions participants proposed. Subsequently, a cross-comparative analysis grouped observations into "male pov" and "female pov" where differences emerged, utilising comparative tables to systematically highlight these divergences. Ultimately, this rigorous method of analysis culminated in the development of a substantive theory: the Digital Violation Discrepancy, which emerged directly from the findings.

3.1.1 Limitations of the Focus Group Approach

Despite careful planning, the focus groups presented several practical limitations that warrant acknowledgement. Recruitment challenges impacted the study's initial design and execution; a lower-than-anticipated response rate meant that the researcher, contrary to initial hopes, facilitated the all-male group; absence of an external male facilitator could have impacted group dynamics or participants' candidness, as suggested by literature on gender-matched facilitation (Benstein Miller, 2015, p.66-67; Kollock et al., 1985, p.42). However, observations during the session suggest that participants remained highly engaged and open, mitigating significant concerns about this limitation.

A more substantial limitation, also derived from recruitment issues, stemmed from the cultural homogeneity of the participants. Despite efforts for diverse recruitment, the persistent shortcomings led to a predominantly homogenous participant pool: of 14 participants, 12 were white-European, of which all female participants were white-European. While some female participants identified as queer, the limited sample size prevented a dedicated intersectional analysis; this lack of intersectional diversity, particularly race, impacts the generalisability and nuance of the findings, especially amid the discourse of deepfake harms (Kitzinger, 1995, p. 300; Krueger & Casey, 2015, p. 121-122). Diverse backgrounds can shape perceptions of technology, privacy, consent, and gender,

suggesting the findings may reflect a white-European feminist perspective. Future research should aim to prioritise recruiting an intersectional participant base to capture the full spectrum of experiences.

3.2 Critical Discourse Design and Implementation

To address RQ3 and RQ4 - further understanding accessibility and how media represents AI-generated deepfake harms, particularly minor involvement – the study employed critical discourse analysis [CDA]. This approach aimed to uncover how these narratives oversimplify, prioritise sensationalism, and, consequently, fail to address systemic factors. This representational failure, in turn, highlights the necessity of a more nuanced and complex societal response than the typical media narratives suggest.

Informed by van Dijk's socio-cognitive approach, CDA was chosen for its utility in examining how discourse not only describes social reality, but also actively constructs and reinforces social representations, ideologies, and power relations within a given context (Van Dijk, 2008, p.6-8; Wodak, 2015, p.64). This method emphasises the interplay between text, cognition, and society, providing a lens to analyse how media discourse shapes public perceptions of deepfake harms. In practice, a socio-cognitive CDA involves systematically probing selected media texts, moving beyond surface-level textual analysis to investigate underlying cognitive models, schemata, and shared knowledge structures influencing how journalists frame events and how readers ultimately interpret them (Van Dijk, 2008, p.6-10). By deconstructing linguistic choices, thematic prominence, and rhetorical strategies, this research revealed how media discourse actively frames the social representations of NCII crimes, specifically examining the portrayal of minors' dual role and the perceived efficacy of various societal responses. This method is well-suited to research, allowing examination of how the "who, what, and why" of deepfake harms are presented, thereby illuminating the broader social and cognitive implications of these media portrayals.

Data Collection

Media text selection followed rigorous criteria, ensuring research relevance and representativeness of mainstream public discourse. Seven articles were selected: two from the United Kingdom (The Guardian, The Independent), three from the United States (NBC News, CNN, CNN), and two from South Korea (The Korea Herald). This deliberate geographical distribution reflected three distinct stages in deepfake acknowledgment and regulatory response. In this hierarchy, the United States is perceived as regressing in AI regulation, evidenced by the removal of the Biden administration's 2023 ethical AI act and a proposed 10-year moratorium on AI regulation (Akselrod & Venzke, 2025, para.1; Hendrix & Lima-Strong, 2025, para.1-2). The United Kingdom appears to acknowledge deepfakes as a serious issue, proposing legislation, yet the pace of implementation remains comparatively slow and there are few punitive measures (Hörnle, 2025, para.1-2). In contrast, South Korea, at the issue’s epicentre, transitioned from a legislative vacuum in August 2024

(see *Appendix A* for further information) to implementing comprehensive, multi-dimensional measures, including punitive provisions for minors as of 2025 (Seok et al., 2025, p.2-10). This country-specific selection allowed for an additional analysis layer beyond a general global take on minor involvement, examining whether differences in country-specific media representations reflected their national deepfake approaches.

Source selection was limited to reputable, mainstream news outlets with broad international recognition. This ensured the analysis accurately reflected public discourse rather than individual opinions, and its potential influence on regulatory approaches. Search efforts focused primarily within news websites. To mitigate potential influence from browser data and cookies, all searches were conducted using a VPN and an incognito browsing window.

Key search terms included: "deepfakes," "minors," "school," "children," and "child." Direct terms like "victim" or "perpetrator" were avoided, as initial exploratory searches suggested explicit victim coverage often overshadowed dual role discussions, and directly using 'perpetrator' could prematurely bias results. Primary articles were identified from The Guardian, The Independent, CNN, NBC News, and The Korea Herald. While other major outlets were also explored, they either did meet the criteria, or lacked the representation of minor involvement needed for this analysis.

Selected articles adhered to the following inclusion criteria: published by mainstream news outlets; between 500-1000 words in length, ensuring sufficient textual content for in-depth analysis; written in English, necessitating the selection of English-language publications even for South Korean contexts; explicitly discussing AIG-deepfake technology or crimes involving NCII; explicitly covering minors' dual role as both victims and perpetrators; and focusing on the harmful or criminal implications of minor involvement. Conversely, excluded texts included: articles that not discussing minor involvement in AIG-deepfake crimes; opinion pieces, academic papers, blog posts, or non-mainstream sources; and articles primarily emphasising technical explanations of the deepfake technology, without reporting on specific incidents or broader societal implications.

Data Analysis

Drawing from van Dijk's framework, the CDA systematically examined the selected articles to identify how discourse constructs and reinforces social representations around minor involvement in NCII crimes. This began with a close reading of each text, noting emergent categories, themes, and concepts. Notes were then cross analysed to see which topics were present across all the articles, setting the structure for the analysis. First, minors framing was analysed, examining how minor victims are portrayed, their degree of agency or passivity, the emotional impact conveyed, and whether their voices are directly present or mediated by others. Similarly, minor perpetrator portrayal was assessed, exploring the complexity with which their involvement is presented, such as the acknowledging factors like digital literacy, online culture, peer influence, or a lack of understanding of consequences, versus simplistic depictions as solely malicious actors. Second, accessibility discourse was investigated, identifying explicit or implicit portrayals of minors' ease or difficulty in

encountering or creating deepfakes. This aspect also noted mentions of user-friendly tools, specific online platforms, and the methods of content creation and dissemination highlighted within the discourse.

Emphasis on or acknowledgement of systemic issues was then examined, determining roles and responsibilities assigned to broader systemic actors in facilitating or preventing these crimes. This analysis assessed whether discourse predominantly focuses on individual actions and sensational details, or delves into the underlying systemic factors that contribute. Next, article tone and linguistic choices were critically assessed, identifying language that might reflect a digital immigrant perspective, such as expressions of shock, a perceived lack of understanding about online behaviours or technological realities commonplace for digital natives, or a tone that suggests a struggle to grasp the evolving nature of digital harm.

Finally, the call for response was analysed, investigating what responses articles explicitly suggest or implicitly call for, and from whom. This analysis also evaluated whether proposed responses are presented as comprehensively addressing the full complexity of the issue or if they appear limited in scope and understanding. Amidst all the layers of analysis, attention was paid to the country-specific differences in representation, and the extent this aligns with each country's AI approach - borrowing from the focus group analysis and employing a constant comparative method. The CDA analysis similarly employed comparative tables to group differing representational approaches ("US only", "UK only", "Western Rep"[7] and "South Korea Rep") and how they align with the countries approaches to regulation.

3.3 Positionality

As the sole researcher for this qualitative study, positionality's role in shaping the research process warrants consideration. When investigating the disproportionate gendered harms of deepfakes, identifying as a female researcher inherently informed the approach taken, fostering a particular sensitivity to women's and girls' experiences with digital violence. This personal engagement stems not only from academic interests in analysing technologies' societal shortcomings - leading to an inherently critical view of technological advancement - but also from my own lived experiences with technological violence.

Grounded in feminist standpoint epistemology, this study explicitly acknowledges knowledge is socially situated. While committed to strong objectivity, the interpretive nature of qualitative inquiry means the analytical lens, shaped by one's own background and critical perspective, inevitably guides theme and pattern identification. Notably, for the CDA analysis, this

[7] The researcher acknowledges that "Western Rep" is a large generalisation to make from only two western countries, and thus references to this category should be understood in under an abstract definition of "Western Representation" opposed to a holistic, country-by-country, synthesis of media representations.

involved a conscious focus on highlighting discrepancies, omissions, and implicit framings within media representations that resonated with personal understandings of gendered power dynamics and technological control. To combat this, reflexivity was maintained throughout the research process, continually acknowledging and critically examining researcher biases and assumptions, considering their influence on data collection, facilitation, and subsequent interpretation. This commitment to reflexivity is central to upholding the principles of strong objectivity within feminist standpoint epistemology.

4. Focus Groups: A Gendered Lens Dividing Deepfake Technology

Grounded in standpoint epistemology, the focus group findings led to the development of an emergent substantive theory: the "Digital Violation Discrepancy." This theory posits that individual experiences, particularly shaped by gender and generation, influence understandings of deepfake harms and their underlying mechanisms. The following subsections detail emergent categories derived from three focus groups - an all-female group [F1, F2, F3], an all-male group [M1, M2, M3, M4], and a mixed-gender group [Male; {CM1, CM2, CM3}, Female; {CF1, CF2, CF3, CF4}] - highlighting similarities and differences across gendered perceptions. Data reveals key themes and patterns in participants' perceptions of AI-generated deepfakes concerning harm, prevention, and minor involvement; drawing upon participants' direct dialogue and observed group dynamics, these findings serve to illustrate and substantiate the core tenets of the Digital Violation Discrepancy.

4.1. Perceived Accessibility and Technical Understanding

All focus groups began by exploring prior knowledge of technical processes and perceived accessibility to establish participants' preliminary understandings and perceptions of deepfake technologies, which were unanimously negative. Across all three groups, individuals reported frequent encounters with [non-sexually explicit] AIG content online, especially via social media platforms, often struggling to distinguish authentic from fabricated media. As F2 expressed, navigating digital platforms has become an exercise in uncertainty: "I'm not sure what's real."

While participants demonstrated a conceptual awareness of deepfakes, technical understandings remained superficial. Perceived ease of creation emerged as a key concern, with participants recognising the increasingly user-friendly nature of deepfake generation. M2's observation that such technologies are "only a few clicks away" highlights this accessibility, suggesting an underlying technological incentive for potential misuse. The male focus group, potentially influenced by M2's technical backgrounds in software engineering, displayed a marginally more nuanced understanding of this accessibility, discussing the ease with which deepfake creation advice could be obtained through online searches or by consulting AI platforms. The mixed group's references to celebrity deepfakes and proliferation of AIG art further underscored the growing normalisation of this kind of synthetic media.

The perceived ease of *encountering* deepfakes, coupled with limited technical understanding of their *creation*, suggests a potential vulnerability rooted in a digital literacy gap, where the development of these technologies is outpacing societies technical understandings.

4.2 Multi-Dimensional and Severe Harm

The focus groups revealed a multidimensional understanding of the harms inflicted by deepfake technologies, exposing how digital manipulations penetrate and destabilise individual lives.

Participants unequivocally recognised the psychological and social consequences of these AIG-NCII. The perceived severity of AIG-NCII harm targeting adults was unanimously viewed across all groups as severe; participants' initial reactions to AIG-NCIIs of adults were overwhelmingly negative, described as "Extremely negative", "pretty horrible", "gross", and "very inappropriate". The identified harms included:

Emotional/Psychological: Deepfakes cause negative emotional impacts, leading victims to feel their "life is ruined", potentially contributing to suicidal ideation. One female participant highlighted significant impacts on "mental health and self-perception", noting how deepfakes "hit you, but at least you've had a certain years of experience in life to have more ways to deal with it" as an adult compared to a child. The female group noted the deep discomfort and emotional toll, while the male group discussed the potential for extreme emotional distress and suicidal ideation. Psychological harm was linked to an inability to consent and a loss of agency over one's own body.

Social/Reputational: Deepfakes can destroy a person's reputation, impacting relationships, employment prospects, and leading to social ostracization or discomfort. Uncertainty of who might have seen the deepfake creates "paranoia" - M4 captured this anxiety, describing the constant uncertainty of wondering whether colleagues, classmates, or community members have encountered the fabricated intimate content. Groups also noted how this social vulnerability extends into professional domains, recognising how such digital violations could impact employment prospects and interpersonal relationships.

Safety: While less frequently discussed, participants identified that deepfakes could lead to safety implications. Specifically, CM3 voiced concern that sexually explicit deepfakes depicting individuals in compromising scenarios, such as “CNC [Consensual Non-Consent] or rapeplay”, could invite violence towards women in real life. This was seen as potentially setting a "dangerous standard" by normalising harmful content, posing a danger that others might interact in real life based on what they have seen in fabricated content.

Erosion of Trust: Participants also recognised a broader epistemological threat: the erosion of digital trust. As F1 noted, deepfake technologies pose a challenge to notions of evidential reliability, potentially impacting judicial systems where video footage can no longer be accepted as truth, and the effects of deepfakes are perceived as "real while sources are fake". This creates a precarious situation for victims, as the potential defence of "oh, it's a deepfake" might be dismissed or not believed.

Notably, participants' physical discomfort during these discussions - fidgeting, avoiding eye contact, speaking in hushed tones - revealed the emotional weight of these technological violations. Harm was consistently framed beyond an abstract technological problem, as a personal assault on individual autonomy and dignity. Moreover, CF1 stressed how the widespread accessibility of deepfake technologies transforms these violations from potential aberrations to personalised threats by articulating the impacts on consent and autonomy:

"...overall loss of autonomy and the idea of consent being maybe less established – not that its super established today – but maybe less, and the idea that you can do whatever you want and see someone naked without their consent – well "naked", you're not really seeing them – but the idea of their nakedness… the idea of consent – the barriers are blurred a little bit more for the people that generate this stuff than they would be, and that it might invite people to view anyone, women or men, but yeah especially women, in a certain way that may be giving them less consent." (CF1)

This perspective suggests that the creation of a single non-consensual deepfake, particularly of a woman, inherently projects a conceptual risk and vulnerability across an entire demographic. This recognition underscores the need to understand deepfakes not as technological curiosities but as mechanisms of digital violence that impact notions of consent, identity, and bodily integrity in the digital age.

4.3 Gendered Perceptions of Harm and Vulnerability

The focus group discussions revealed significant gender differences in perceiving and understanding the deepfake harms. A strong consensus emerged among female participants, that deepfakes disproportionately affect women, with CF2 stating, "technically it could happen to anyone, but I assume it probably disproportionately affects women – especially famous women as well"; a perception substantiated by statistical evidence showing that approximately 99% of deepfake pornographic content targets women (Jacobsen & Simpson, 2024, p.91).

Female participants showed a nuanced understanding of deepfake harm extending beyond technological violation, consistently linking these technologies to broader systemic issues of online harassment and sexualisation of women's bodies. Discussions around consent were particularly revealing for the female perspective; F1 emphasised, "No consent is ever given, and it's like literally the most baseline thing when you're, like, indulging in sexual encounters. Consent is literally the most important thing". This frames the creation of AIG-NCII as a fundamental violation of personal autonomy, understanding deepfakes as a form of sexual violence that fundamentally challenges bodily integrity.

Conversely, male participants, while acknowledging the harm, approached the issue differently. M4 connected deepfakes to "higher rates of just misogyny and this idea that women are only seen as sex objects", demonstrating awareness of the broader systemic issues. However, their analysis often remained surface-level, tending to focus on technological mechanics or potential perpetrator motivations rather than personal impact on victims. Crucially, the male group acknowledged the gendered disparity in social repercussions, in that women face harsher stigmas and consequences than men for similar content. M2 stated it is "probably even worse... if it's a woman, because... the woman will have a lot of stigma attached to her if she's seen doing stuff online". M1 noted that men often don't face the same repercussions for leaked content, stemming from a

"privilege that comes with being a man", while acknowledging that the sexualisation of women's bodies makes deepfakes "always harmful" for them.

When exploring scenarios involving male victims, participants noted similarities in types of harm – emotional distress, social consequences, and reputational damage. However, crucial differences emerged in the perceived severity. The female group reinforced that the lack of consent is the intrinsic issue and origin of harm for *anyone*, regardless of gender, although, while the initial impacts might be similar to that of a female victim, they did expected any long-lasting results to be less harmful for a male victim. The male participants, on the other hand, often introduced the caveat that severity of harm for male victims "depends on the content" - a significant divergence from the female group's view. M2 suggested that if a deepfake of himself showed something he deemed "funny" or could be "proud of," like having a "massive penis," he might find it less harmful, or even positive. This contrasts the experience of female victims and highlights a potential gendered double standard in perceived harm, linked by M1 to the "masculinity image".

A notable moment occurred when CM3 suggested women would have a "more emotional process…" because "...statistically, we have figured out that women encounter more emotions than men". This comment, drawing visible discomfort from female participants, exposed problematic stereotypical assumptions about gender and emotional experience, revealing ingrained biases that often undermine serious discussions about technological violence.

The group dynamics themselves were revealing. The female group's body language – closed off, self-comforting – during discussions of sexual deepfakes contrasted with their otherwise proactive participation. The male group's visceral reactions suggested an underlying recognition of the technology's potential for harm, even if their analytical approach differed. A key observation from the male discussions was their tendency to conflate deepfakes with the leaking of real images, such as nudes or revenge porn, suggesting difficulty grasping the unique violation of *fabricating* content. Furthermore, the male participants showed an inclination towards dismissing or downplaying the severity of sexual harms caused by deepfakes, instead favouring political harms, scams, or other non-sexual crimes enabled by the technology. This led to more in depth answers when understanding the perpetrator (e.g., motives for political deepfakes or scams) compared to understanding the victim of sexual harm.

The focus group data thus reveals a significantly divergent gendered perception of deepfake harm, directly addressing RQ2 by illustrating how societal roles and power structures shape interpretations of digital threats. These patterns reflect the core tenets of the Digital Violation Discrepancy theory, demonstrating how individual experiences, particularly shaped by gender, influence understandings of deepfake harms.

Female participants consistently highlighted deepfakes as a disproportionately gendered harm, linking the technology to broader systemic issues of online harassment and the pervasive sexualisation of women's bodies. Their discussions underscored that for women, deepfakes are not

merely technological anomalies but a continuation of deeply entrenched subjugation, aligning with Plant's (1997) conception of digital tools as contemporary extensions of historical gender-based control (p.38-41). Crucially, female participants foregrounded consent as the foundational violation, viewing the act of generation itself as an inherent ethical breach regardless of whether the content was ever distributed - directly echoing Butler's (1999) theory of gender performativity, where deepfakes hijack the performed construction of gender by imposing false acts onto an individual's digital likeness, thereby attacking their bodily autonomy (p.29-31). The female group's insistence that "no ethical intention" (F2) could exist behind such creation underscores their focus on the inherent violation of digital agency from inception. These findings support the thesis's core argument that deepfakes function as instruments of technological sexual violence, exposing how digital technologies become elaborate mechanisms for perpetuating gender-based control, with differing perceptions highlighting how technology is weaponised within fluid digital networks to target women's performatively constructed identities.

Conversely, the male participants' approach, often more perpetrator-focused and less centred on victim experience, reflects broader androcentric biases in understanding technological harm, aligning with virtual feminists' critique of how technological violence is conceptualised and addressed. Ultimately, the findings thus far place deepfakes at an intersection of gender, power, and digital manipulation, where existing patriarchal structures find new, algorithmically sophisticated methods of control and violation.

4.4 Contested Source of Harm: Generation vs Distribution

Questions regarding generation and distribution of deepfakes revealed another gendered juxtaposition in conceptualising deepfake harm, particularly concerning the moment of violation. A notable point of contention emerged around whether harm is primarily located in the generation or distribution of AIG-NCII.

Female participants consistently positioned consent as the foundational issue - F2 explicitly argued creation itself constitutes a violation, stating that there is "no ethical intention possible" behind generating such content. This view suggests that the mere act of manipulating someone's likeness without consent represents a fundamental breach of personal autonomy, regardless of dissemination. While the female group agreed that distribution undeniably amplifies the harm due to its wider reach and impact, they maintained generation is where the harm originates. F2 linked distribution back to the initial creation, explaining that seeing deepfakes shared online gives others the incentive and perceived "power" to create their own, suggesting a cycle where distribution encourages more generation; that without the generational harm, the distribution harms would be a non-issue.

In contrast, male participants predominantly viewed distribution as the primary source of harm, prioritising visible, external, or publicly observable consequences over the violation of non-

consensual creation itself. M4 suggested privately created deepfakes cause minimal damage, as “the person you generate might not even know you did that", contrasting with distribution where “that's when it becomes a legal problem”. This suggests a focus on harm with tangible, legal, or publicly evident consequences, potentially minimising psychological and ethical violence inherent in the act of AIG-NCII, such as the loss of agency over one's own body and likeness or the anxiety stemming from the uncertainty of not knowing if or how one's manipulated image exists or might spread. M2 continued, noting that creation often leads to distribution because "that's just the way people do stuff" - they find it amusing and inevitably share. This view, while acknowledging the connection between creation and sharing, still frames the primary problem around the act of sharing as the point where the content becomes uncontrollable and widespread.

The mixed group highlighted this dynamic. CM2 initially suggested harm was "both in hand", but subsequent discussion revealed a tendency to prioritise distribution, particularly among male participants. CM3 claimed distribution was the main harm, aligning with the male focus group consensus, to which CF2 reframed the issue, describing deepfake generation as "exercising a form of power over another person that does not consent to being used", applying this principle not just to sexual content but also to political deepfakes, highlighting that using someone's voice or image to say something they didn't consent to is automatically harmful. This aligns strongly with the female group's view that generation is where the harm begins due to rudimentary consent violation.

This split view reflects deeper societal dynamics around consent and bodily autonomy. The female participants' emphasis on generation as the primary violation aligns closely with feminist perspectives that recognise harm extends beyond physical manifestation. Their view suggests that the act of creating NCII is itself a form of technological violence that fundamentally challenges an individual's right to control their own image and identity. The male perspective, by contrast, often minimised this ethical breach of creation and consent violation, focusing instead on the potential social consequences of distribution, suggesting a societal tendency to overlook non-physical forms of violation, particularly those involving digital spaces.

Critically, when discussing scenarios involving minors, this divide became more pronounced. Female participants continued to unanimously view generation as inherently harmful due to the complete lack of consent. F2 articulated this point, stating that unlike with adults - where there could hypothetically be a discussion around consent and mutually agreed-upon creation - there is "not a single instance" that makes it ethically acceptable to create sexual content involving a minor. This strongly reinforced the female group's pre-existing priority on the generative act as the harm's origin, forming an even clearer consensus when minors were involved. Male participants, however, continued to emphasise distribution as the primary concern, although with more divided opinions on the matter. Thus, while the male group did not reach the same level of unanimous consensus as the female group on the primacy of generational harm for minors, the discussion around minors did introduce a stronger acknowledgement among some male participants that the generative

act itself carries significant risks, particularly in normalising disturbing interests and potentially leading to real-world harm, such as paedophilia.

This distinct opposition in perception of the source of deepfake harm - generation versus distribution - further substantiates the Digital Violation Discrepancy theory, illustrating how entrenched gendered perspectives shape interpretations of digital violence. Female participants' emphasis on generation as the primary violation aligns with their previous perspective of creation and consent violation as the origin of harm, and with feminist perspectives that recognise harm beyond physical manifestations, highlighting how the act of creating NCII challenges an individual's right to control their own image and identity. Juxtaposing this is the male tendency to minimise the ethical breach of creation, focusing instead on social consequences of distribution. While acknowledging deepfake's capacity for harm, the male focus gravitated towards the tangible, publicly observable consequences, such as legal problems or reputational damage, rather than the psychological and ethical violence involved with NCII itself. When contemplating male targets, male participants qualified the severity of harm based on the content of the deepfake, with some even suggesting potentially "funny" or "proud" outcomes, contrasting the female perspective, which views any non-consensual manipulation as inherently harmful, regardless of content.

This gendered double standard not only minimises the impact on male victims of sexual crimes but also exposes a societal discomfort with acknowledging non-physical forms of sexual violation; the male tendency to assess harm based on content, especially in relation to a "masculinity image", often leads to a downplaying of the severity of non-consensual acts when applied to men (Langdridge et al., 2023, p.1-2). This perspective risks undermining the basic violations of bodily autonomy and dignity inherent in *any* non-consensual manipulation, regardless of victim gender or the content nature. Male participants also tended to conflate deepfakes with the leaking of real images or to pivot discussions towards political harms and scams, further revealing a potential struggle to grasp the unique nature of fabricated content and a broader societal bias that often overlooks sexual violence in digital spaces (NSCRC, 2025, para.1-2; Hlavka, 2016, p.22-26). These patterns in perception reinforce the thesis's overall argument that deepfakes operate as instruments of technological sexual violence, perpetuating gender-based control within fluid digital networks, with this split reinforcing how deepfakes, as instruments of technological control, are conceptualised differently based on gender, and the need for a more nuanced understanding of consent and autonomy in the digital age.

4.5 Minors as Heightened Victims and Perpetrators

Addressing RQ3, focus group discussions revealed how understandings of minors' involvement in deepfake crimes reflect the reality of technological accessibility and their unique vulnerabilities. All focus groups concurred that deepfakes targeting minors represent a uniquely devastating form of harm that is worse than for adult victims. Minors' developmental vulnerabilities

emerged as a focal point. Participants highlighted immature cognitive and emotional frameworks that make young people more susceptible to deepfake harm. F3 noted that the underdeveloped prefrontal cortex - responsible for understanding long-term consequences - creates a storm of vulnerability and recklessness among minors involved as both victims and perpetrators. M4 similarly highlighted this, noting that a 15-year-old is particularly vulnerable to psychological consequences, potentially even suicidal ideation. Critically, discussions linked this heightened vulnerability to a digital environment where deepfake technologies are perceived as increasingly accessible, aligning with their actual widespread availability (Meyer, 2025, para. 2-5). Participants recognised that deepfake creation is "only a few clicks away", contributing to a pervasive sense of uncertainty about distinguishing authentic from fabricated content. This digital literacy gap, where technological advancement outpaces public understanding, thus leaves minors highly susceptible; their substantial online footprints, coupled with the ease of user-friendly deepfake tools, provide abundant, often unknowingly contributed, source material.

Furthermore, while agreeing on developmental shortcomings, discussions revealed gendered differences in explaining the underlying reasons *behind* minor involvement. Female participants emphasised systemic issues like normalised online sharing of minors and exposure to harmful content, whereas male participants leaned more towards explanations rooted in inherent traits of youth and a general lack of understanding ("idiots," "kids are stupid"). These gendered responses and perceptions highlight that, while developmental vulnerability is universal for minors, its experience and understanding in deepfake harms are impacted by gender and societal factors.

At the revelation of minor involvement in deepfake crimes, participants were shocked and saddened by the scale, but ultimately unsurprised, reflecting a nuanced understanding of contemporary digital culture and accessibility of technology. M4 captured this sentiment, describing how nine-year-olds today are exposed to content far beyond what previous generations experienced, highlighting the rapid transformation of online spaces. This paradoxical dual role highlights the complicated realities of Prensky's (2001, p.1) digital native generation, with their innate comfort with digital manipulation meaning that deepfake creation can be viewed as a casual extension of online activities, rather than a severe ethical breach (Y.Choi, 2024b, para.7). Participants identified multiple interconnected factors contributing to minors' dual role. The normalisation of harmful content emerged as a primary theme – growing up with readily available sexual content on platforms like X and Reddit makes accessing and engaging with NCII material seem normal to minors who "don't know better yet". F1 described this as a form of "digital grooming" that normalises harmful behaviour and could encourage creation. CM2 observed that growing up surrounded by such technologies shapes foundational perceptions of what is acceptable, creating a distorted understanding of reality and ethical boundaries in digital spaces. Moreover, peer influence and online validation were factors in motivating minor perpetrators. M2 noted how digital anonymity allows people to "do stupid things," an observation capturing online behaviour dynamics, where anonymity

provided by encrypted chat groups enables environments where harmful behaviours could be normalised and executed with a sense of impunity.

A significant concern, raised by both genders, was the normalisation of posting children online. CF1 highlighted a vulnerability with parents posting images of young children or infants on social media providing source material for deepfakes; participants saw this trend as contributing to a broader erosion of understanding around children's digital consent. While the high statistic of minor perpetrators was central, participants also found it very concerning that a not insignificant portion (27%) were still adults. This concern about adult perpetrators was linked to the fear that AIG-CSAM could normalise or even encourage paedophilic behaviours. M2, M4, F1, and F3, each discussed the risk that allowing explicit material of children, even if "not real", could lead people to grow accustomed to it, normalise it in their minds, and potentially lead to actual interest in real children or taking "generation into action" in the real world. CF2 noted that generating AIG-CSAM requires "criminal and creepy energy" and that creation itself, regardless of distribution or the minor's awareness, is always harmful. This contrasts slightly with some initial male perspectives that focused more on distribution as the main harm for adults, though this view shifted when discussing minors.

The dual role of minors as both victims and perpetrators emerged as particularly concerning. CF1 highlighted how minors "churning out this content" increases the likelihood of it reaching broader audiences, while CM3 feared a cyclical pattern: more minor perpetrators inevitably leading to more minor victims. This observation revealed a concerning technological ecosystem where accessibility and online culture facilitate a self-perpetuating cycle of harm. Critically, participants recognised that adults - characterised as digital immigrants - struggle to comprehend the nuanced online environment that shapes minor's experiences. The disconnect between generational technological understandings creates significant challenges in addressing and preventing deepfake-related harms.

This analysis highlighted that minors are not merely collateral damage but primary targets in this emerging technological landscape. Their immersion in digital culture, combined with developmental factors like limited understanding of consequences and a search for online validation, positions them as both vulnerable victims and potential perpetrators. These findings expose an intricate technological ecosystem where traditional notions of victimhood and agency become increasingly blurred, with the deepfakes demanding a nuanced, multi-stakeholder approach to understanding and prevention against *all* ages of perpetrators.

It is important to acknowledge that the focus group participants, all under 30, by definition fall within the digital native category, having grown up immersed in online spaces - this shared generational context likely influenced their nuanced understanding of deepfake accessibility and the factors contributing to minors' involvement as both victims and perpetrators. Their lived experience within these digital environments afforded them insights into the normalisation of certain online behaviours that might be less apparent to digital immigrants who adopted these technologies later in

life (Prensky, 2001, p.2; Harding, 2016, p.138-140). Consequently, perceptions of minor involvement and the severity of deepfake harms may differ considerably among older demographics; future research could productively explore these questions with participants aged 40 and above to gain a more comprehensive, generationally diverse understanding of this Digital Violation Discrepancy. However, ultimately, the focus groups revealed that participant perceptions regarding minors' dual role and the accessibility of deepfake technologies do largely align with the actual technological landscape, setting the stage for the CDA to further explore how media representations further shape these understandings.

4.6 Responsibility and Prevention

Focus group discussions around deepfake responsibility and prevention s emphasised the challenges of mitigating deepfake harms in an increasingly digital world. Participants universally found that addressing deepfake risks requires a multi-stakeholder approach, with responsibility distributed across parents, educational institutions, technology companies, and governmental bodies:

Parents: Male participants strongly emphasised the responsibility of parents. CM2 stated that parents "don't do their jobs", and that digital education primarily "happens at home". This perspective came with increased parental control, such as restricting Internet access, monitoring browsing history, or delaying smartphone use for young children. While the need to educate parents on *how* to monitor was mentioned, focus was on parents implementing controls and taking direct charge of their child's accessibility, expressing a sentiment that parents are negligent if they know the dangers and do not act. This perspective was not unanimous. Female participants, while agreeing that parents should play a role, expressed reservations about relying on parents, highlighting that many parents may lack sufficient knowledge or experience with rapidly evolving AI and deepfake technologies. Female participants were also concerned about the inconsistency of parental education, noting difficulty ensuring consistent and equal education considering issues like "child neglect" - CF3 further probed "...but also what do you do when they are not even any parents?". This clash highlights the divergence in viewing the problem: individual parental oversight (more prevalent in the male perspective) versus a systemic issue requiring broader educational and regulatory frameworks beyond the home (more strongly articulated by the female participants).

School: Female participants advocated for schools taking a leading role in this education, with F1 arguing that educational institutions are the "foremost teachers," possessing more comprehensive knowledge than a parent. Female participants envisioned holistic approaches to digital literacy education extending beyond technical skills, focusing on ethical considerations, societal impacts, and critical understanding of online environments, with an emphasis on teaching young people how to critically navigate the internet, rather than simply providing technical instructions. The female group framed this as being introduced as rudimentary as "sharing is caring". Men, while agreeing educators should play a role, raised concerns about overburdening teachers with

this additional responsibility. Educative measures from the men regarding deepfakes were framed akin to sex-ed classes in school – introduced later (early adolescence) with less emphasis around ingraining it as a core philosophy, and more in relation to sexual acts and consent.

Tech companies: Female participants placed significant responsibility on technology companies, advocating for prevention at the source (generation). Arguing that platforms "shouldn't even be able to make it" possible to create such content, they pointed to existing efforts by some companies in preventing the use of known people's images - though acknowledging that people would try to bypass these safeguards easily. This perspective aligns with the female group's broader emphasis on generation as the starting point of harm due to the violation of consent. There was also a sense of scepticism about fully trusting companies to regulate themselves, and the necessity of complimentary external governance. Male participants focused more on practical difficulties of implementation and circumvention. M2 suggested restricting deepfakes to research purposes but quickly pivoted to the ease of bypassing such rules with VPNs if regulation is only country specific. M1 noted that while efforts are made by companies, circumvention is always possible; CM1 argued that "most of the blame lies with private AI companies that make this so incredibly accessible with no regulation," especially for AIG-CSAM. Some male participants suggested regulating or even banning AI companies that don't prevent creation, but this was often paired with the acknowledgement that tech-savvy individuals could bypass safeguards.

Legal/Governmental: Both male and female participants saw governmental bodies as key actors, sharing an understanding that laws are needed, particularly for adults, serving as preventative measures and setting societal expectations. The all-female group placed a stronger emphasis on governmental regulation targeting the creation (generation) of harmful content, while female participants in the mixed group also pivoted towards the need for governmental bodies to set ethical limits on both generation and distribution. While constantly acknowledging the difficulty in defining these limits, this focus on restricting creation at the source reflects the female group's broader emphasis on consent violation as the primary harm. Female participants also noted the need to address networks, like those on Telegram, and look domestically at the issue, not just internationally. For minors, the women strongly favoured rehabilitation and education over legal punishment. Male participants, however, often focused on practical challenges of implementing and enforcing legal and governmental controls, highlighting how individuals could easily circumvent rules using VPNs, the dark web, or the Hydra-like nature of online networks where shutting one platform down leads to others emerging. Their discussions often shifted focus to other types of harm, like political deepfakes or scams, when discussing the dangers, implying a different area of priority for regulation compared to the sexual deepfakes emphasised by the female participants. M2 also brought up the complexity of regulating internationally due to differing national cultural standards and ages of consent. Their perspective was summarised by M1's claim that "So even if you do put these very strict things, people that like it will always find a way"

Overall, female participants viewed responsibility and prevention through a systemic lens, strongly advocating for rehabilitation and education over punishment, particularly for minors, and focusing intervention at the source by preventing the creation of harmful content, summarised by the sentiment that one "shouldn't even be able to make it". In contrast, the male discussion, despite their demographic forming the largest group of perpetrators, and some participants noting a potential for "pride" or less negative consequences for male targets, often appeared more complacent, emphasising the practical challenges of implementation, downplaying the harm of creation if kept private, shifting responsibility (often to parents), and making excuses as to why robust protective measures seemed difficult or unnecessary in certain contexts due to the inevitability of circumvention or the nature of "people will be people".

These discussions exposed yet another gendered divide in understanding technological risks; participants' discussions reflected what Plant (1997, p.177) might describe as the fluid, interconnected nature of digital networks, where traditional boundaries of control become increasingly blurred. The challenges of prevention mirror the theoretical framework's understanding of technology as a complex site of power negotiation. Ultimately, the focus group data revealed prevention as a multifaceted challenge that cannot be addressed through a single approach - addressing deepfake technologies demands a multi-layered response that recognises the intricate ways technological harm is generated, distributed, and experienced across different social groups and generations.

5. Critical Discourse Analysis: Media's Framing of Minor Involvement

Building upon chapter 4, the following presents the critical discourse analysis findings of media representations of AIG-NCII crimes involving minors as both victims and perpetrators. Informed by van Dijk's (2008) CDA approach, the analysis draws from seven mainstream news articles across three distinct geographical regions - the US (perceived as regressing in AI regulation), the UK (acknowledging deepfakes but slow in implementation), and South Korea (proactively implementing comprehensive measures) - with this national hierarchy serving as an underlying analytical lens. Moving beyond descriptive content, this analysis critically examines how media discourse actively frames deepfake harms, particularly concerning minor involvement. Specifically, it explores how these narratives often inadvertently reflect or reinforce the informatics of domination (Haraway, 2016, p.29-33) by focusing on observable harms while obscuring underlying mechanisms of technological control and the deeper societal tensions that Plant (1997) identifies. The media routinely present these incidents as a "severe" and "urgent crisis" and often emphasise consent violation ("without the consent of the girls"). Media representations also consistently portray the technology as increasingly advanced and accessible for distribution, even if creation's ease is more often implied than explicitly detailed. Thematic subsections will centre on this critical analysis, integrating selected textual evidence to illustrate key discursive patterns regarding accessibility, minor's unique vulnerabilities, gendered impacts, and calls for systemic intervention.

5.1 Understanding Minor's Involvement and Accessibility in Media Representations

The media's portrayal of deepfake accessibility for encountering and distribution offers a partial, yet telling, reflection of the technology's actual reach and ease of use, simultaneously highlighting a disconnect in societal understanding of said accessibility. Articles consistently emphasise high accessibility of deepfake content for encountering and distribution, detailing its rapid spread across social media platforms and encrypted chatrooms. Images are repeatedly described as "circulated online" and widely shared on "social media platforms" like Snapchat and through "Telegram group chat rooms," with media noting content "continue[s] to quickly spread online despite company policies banning such content". This constant presence in minors' digital environments implicitly shapes an understanding that deepfakes are an unavoidable part of modern online life - an observation resonating strongly with focus group findings, where participants described the "normalisation of harmful content" in online spaces as a form of "digital grooming". Such normalisation for minors aligns with Prensky's (2001) digital native, whose innate fluency with networked environments can lead to a desensitisation to online risks, facilitating a "perceived ease of creation" as a "critical concern". This ease of widespread dissemination also speaks to the fluid and pervasive nature of digital power within the informatics of domination, where content, once released, becomes nearly impossible to contain.

However, the media's representation of deepfake creation accessibility is notably less explicit. The frequent involvement of male middle school students or teenagers as alleged perpetrators - generically labelled as "teenager" or "boy" with portrayal largely focused on alleged actions (producing, possessing, distributing) - combined with emphasis on the scale (“30 schoolgirls” “around 50 girls”) implies a low technical barrier to enter. Similarly, the common use of "photos posted to social media" as source material represents readily accessible components for deepfake generation. Moreover, while NBC (Source 1) explicitly states that AI deepfake technology is "becoming more widely available" and mentions "AI-generated images and 'undressing' apps" "available for free on app stores", most articles refrain from detailing the step-by-step process or the specific user-friendly interfaces. This gap suggests a societal understanding that minors can create these deepfakes, but without fully grasping how they do so.

This vagueness in understanding deepfake creation can be attributed to the digital accent identified, primarily in UK/US outlets. Here, a digital immigrant perspective in media may comprehend the outcome (minors creating harmful content) but struggle to articulate the process (the ease of mastering new digital interfaces, the prevalence of online subcultures, or the readily available tools common to the digital natives). This mirrors Prensky’s (2001, p.2) digital immigrants, who, having adopted technology later in life, may understand its outcomes but struggle with the nuanced processes and inherent fluencies of digital natives who effortlessly navigate such tools. This epistemic gap in media coverage, reflecting broader societal digital accents, ultimately limits public understanding of critical entry points for perpetration. This accent manifests as a framing of the technology itself as novel and challenging, with institutions explicitly described as "struggle[ing] to catch up" to these "new technologies". Articles emphasise AI technologies inherent "complexity" and "scary", and deepfakes' "sophistication" and "realism", underscoring a potential lag in understanding and response from established systems less inherently familiar with the digital landscape.

This struggle to comprehend the digital sphere often manifests through explicit expressions of shock and alarm. Adult reactions to deepfake content are communicated with visceral, emotive language - parent describing their daughter "throwing up" after seeing the images or feeling "sick to her stomach" herself. The issue is frequently framed as a "growing crisis", an "alarming rise", or a "wake up call", suggesting a surprise at the scale and nature of these digitally mediated harms. This digital accent of alarm also reflects the challenge of traditional systems to adapt to the fluid and boundary-blurring nature of digital networks that Plant (1997, p.20) describes, where control becomes increasingly difficult. Furthermore, a pervasive anxiety about the uncontrollable nature of digital content is palpable; phrases "You don't know where that is once it's transmitted, when it's going to come back and haunt the young girl", reflect a tangible fear about the permanence and uncontrollable spread of digital content online. This concern is further amplified by frustrations that images "continue to quickly spread online" despite platform bans, highlighting a struggle for traditional systems to keep pace with the digital reality. A perceived disconnect between online

behaviours and traditional societal norms - articulated through language that frames harmful online actions as "out of step with community expectations" or points to a failure in teaching appropriate "standards of behaviour" - suggests a viewpoint that observes digital behaviours with concern, framing them as deviant from established societal norms, and implicitly acknowledging the operation of the informatics of domination through clear codification and enforcement of digital transgressions.

Limited awareness of the technical mechanisms, despite the clear, present danger, restricts a comprehensive public understanding of entry points for perpetration. An interesting find was the use of casual and avoidant language in US/UK articles when referring to perpetrators and their crimes, often prefacing crimes with "allegedly", even for verifiable actions like circulation. This general lack of emphasis on the inherent wrongfulness of the act itself is largely implicit in Western narratives, especially when compared to South Korea. In contrast, South Korean media's representation, which, while not always detailing the precise technicalities of deepfake creation, explicitly identifies and frames them directly through the lens of their crimes, actions, and consequences by detailing specific charges and disciplinary actions - being "booked by the police" or "transferred" and "suspended". Their ‘this is a crime’ approach is consistently communicated through high-volume, explicit reporting that foregrounds minor perpetration, encouraging public discourse with a different level of understanding and emphasis on the acts’ severity. Source 6 (Korean Herald) further substantiates this by providing national statistics, underscoring that a significant proportion of digital sex offenders are minors. The explicit and statistical framing of minor perpetrators in South Korean media establishes them as clear agents of harm, making their involvement a central, undeniable aspect of public discourse.

This analytical gap in Western media (US/UK), potentially stemming from a digital immigrant perspective, reflects a larger challenge in grasping the subtle, pervasive ways that digital technologies - the very platforms of Haraway's (2016, p.29-33) informatics of domination - exert power. Media's struggle to fully articulate the ease and methods of deepfake creation, despite acknowledging its widespread impact highlights a lack of comprehensive interpretation of the forces at play that extend beyond current societal understanding - this struggle in cognisance, reflecting a digital immigrant perspective, which, in turn, potentially underlies the lack of public outrage that might otherwise drive more proactive legislative action (K. Williams, 2024a, para.2-6; Hsu, 2023, para 2-5; Hörnle, 2025, para.1-2 & 8-9).

Furthermore, the consistent lack of exploration and understanding of motivations behind minor perpetration in media representations further impacts the understanding of minors' involvement, appearing unanimously across the articles. Despite their perpetrator focus, even the South Korean articles lacked deep exploration into the underlying factors contributing to minor perpetration, beyond broadly linking behaviour to "broader cultural problem[s]" or the influence of "material on the internet”. Some articles (CNN Source 4, The Guardian Source 5) link deepfake

incidents to broader societal issues like "gender-based violence," a "national crisis," and "misogynistic" behaviour, but fail to probe further, often employing avoidant language when discussing the perpetrator and their crimes (“allegedly circulated”) and prioritising the victim’s experience.

By focusing primarily on the what (actions) and the consequences (institutional responses), rather than the why (complex behavioural drivers), media narratives risk reducing minors to simplistic malicious actors - superficiality that hinders a holistic societal understanding factors encouraging young perpetrators, thereby complicating the development of targeted preventative strategies. This gap in understanding seen in the media stands in contrast to the focus group discussions, which, while not experts, offered granular insights into contributing factors such as the "normalisation of harmful content," "peer influence," "online validation," and the "perceived anonymity of the internet".

By failing to delve into these psychological and social drivers, media narratives fail to explore how such behaviours might stem from underlying societal tensions that need to be addressed in tandem to truly curb the issue. Specifically, they overlook the possibility that these acts are manifestations of male resistance within Plant's (1997) "genderquake" - a radical reordering of gender relations where technological control, exemplified by the creation of "simulations of the feminine" or "digital dreamgirls who cannot answer back," becomes a means to reassert dominance amidst perceived challenges to masculinity (Plant, 1997, p.14-15, p.33, 80). This analytical superficiality means the media fails to fully conceptualise deepfakes within the broader architecture of exploitation that deepfakes represent, missing how these digital tools are integral to perpetuating systemic gender-based control. Consequently, by not exploring these deeper societal and psychological drivers, the media limits public understanding of the systemic issues at play - a point also made by focus group participants who perceived adults as "struggle[ing] to comprehend the nuanced online environment".

5.2 The Need for a More Complex Response to Minors

In response to RQ4, media representations of AI-generated deepfake crimes portray a complex and multifaceted relationship with minors as both victims and perpetrators, revealing an urgent need for nuanced responses. Articles generally frame minors as vulnerable victims through their trauma and impact, strongly emphasising the unique vulnerabilities of minors who are victims of deepfakes. Representations repeatedly ties victim identity to their age group and gender ("teen girls," "middle school students," "schoolgirls, aged 14-18”) and the context, and emphasising potential for long-lasting harm by describing how deepfakes could "come back and haunt the young girl" or the ongoing fear of images resurfacing. The impact on their ability to navigate typical developmental stages and environments (like attending school and classes) is explicitly noted, highlighting how trauma interferes with normal life, while also able to identify a lack of effective "pathways to

recourse" for minors within existing systems, positioning them as vulnerable individuals reliant on external institutional responses. This pervasive victim-centric discourse, particularly highlighting the violation of consent, resonates with Butler's (1999 p.29-31) concept of gender performativity, where deepfakes are understood as a violent imposition of a false digital act, fundamentally challenging the autonomy and bodily integrity of the female self.

However, the manner in which this victimisation and agency are articulated, however, varies significantly between Western (US/UK) and South Korean media, reflecting their differing stages of regulatory development and cultural approaches to AI governance. Western media [US/UK] frames minors as vulnerable victims enduring severe trauma, with a consistent emphasis across articles, featuring specific minor victims (e.g. Francesca Mani and Elliston Berry) as active advocates for legislative change, speaking out and engaging with politicians and the media. In contrast, South Korean media tends to present a more aggregate, statistical view of victims, focusing on institutional responses rather than individual agency post-reporting.

Juxtaposing this, the portrayal of minor perpetrators presents another analytical gap actively impeding the creation of comprehensive response strategies. As aforementioned, the cautious, generic language ("teenager," "boy," "classmate") and frequent use of "allegedly" in US/UK articles when referring to perpetrators reflects discomfort with assigning explicit culpability to minors. This linguistic avoidance, coupled with the media's analytical superficiality regarding perpetrator motivations, fosters a discourse that, ultimately, undermines opportunities for robust preventative or rehabilitative strategies. By failing to explore why young individuals engage in these acts, the media misses key theoretical insights into the intricate societal tensions at play that need addressing. Specifically, they neglect the possibility that such behaviours manifest as a continuation of deeply entrenched female objectification, leveraging digital tools for 'algorithmic sexual punishment' aimed at degradation and humiliation, thereby embodying a continuation of historical mechanisms of control in digitally mediated forms (Plant, 1997, p.38-41). This analytical gap in understanding the perpetrator's motivations and the systemic nature of such digital violence prevents a comprehensive grasp of how deepfakes operate as instruments of power within the informatics of domination.

This disjunction between the problem's depth and its surface-level portrayal stands as a clear reflection of the active approach in legislation and proposed preventative measures. Legal systems, government, and legislators emerged as prominent actors, consistently called upon to take responsibility across the majority of articles. These bodies are explicitly assigned the responsibility for creating and enforcing adequate laws around deepfakes, stemming from a pervasive narrative of systemic failure, where existing legal frameworks are repeatedly highlighted as inadequate to address the specific nature and scale of AI-generated harm. US media (NBC Source 1, CNN Sources 3, 4) heavily emphasises this, explicitly stating that for victims, there are "few, if any, pathways to recourse" in most states and that current state laws "fall short of punishing the content creators". This consistent and urgent focus reflects a perspective of legislative intervention as the primary - and

often only - means to "catch up" to the rapidly evolving technological threat and fill existing regulatory gaps.

One notable finding was the media's overall focus on distribution-based harms over generation-based harms. While articles detail the traumatic impact of images circulating, no articles are explicit about, or argue for, the initial act of violation and harm at the point of creation. This was particularly evident in US/UK media primarily centring on distribution-based harms, emphasising the visible impact of content circulating and advocating for its containment. Potentially, this is explained by the narrative choice to focus on victim experiences – deepfake crimes are creating a complex situation where the victims typically become aware of the crime only upon distribution. However, this narrative choice implicitly downplays the foundational harm inherent in the act of non-consensual creation itself. This representation, it can be argued, translates into proposed laws focused on preventing its dissemination of AIG-NCII content, evidenced by regulations such as the Take It Down Act [USA] or the Online Safety Act [UK], both centred on punishing the act or threat of distribution (Mullin, 2025, para. 1-2; Narayanan, 2024, para. 13-14). This approach, while necessary for victim recourse, reveals a critical disconnect from the foundational act of non-consensual creation itself - the primary violation of bodily autonomy. By focusing on mitigating distribution, these narratives inadvertently downplay the initial violation, failing to fully grasp how deepfakes commandeer identity construction by imposing false performative acts, reflected in solutions that aim to contain the damage rather than address the origins of the harm (Butler, 1990, p.29-31; Mullin, 2025, para. 1-2; Narayanan, 2024, para. 13-14).

Moreover, the digital accent evident in UK/US articles - reflecting a digital immigrant perspective struggling to grasp the complexities of deepfakes and online behaviours – can be seen to translate into an impeded ability to legislate effectively for unseen digital acts like deepfake creation. This analytical gap results in proposed responses that are largely reactive, as aforementioned, focusing on external controls (legislation, platform bans), and struggle against the inherent anonymity of digital perpetration, particularly at the creation stage. Consequently, this reveals a systemic flaw in applying traditional legal philosophies and surveillance models, designed for visible, traceable acts, to deepfake crimes where power is exerted through complex and often opaque digital systems with afforded anonymity. This epitomises Haraway's (2016, pp.29-33) informatics of domination, where power operates through codes and networks that digital immigrant legislative frameworks struggle to fully penetrate or regulate at the point of origin, ultimately resulting in frameworks that attempt to manage consequences without truly comprehending or effectively preventing the underlying digital act itself.

In contrast to Western media's approach, South Korean media explicitly addresses the multifaceted reality of deepfake crimes involving minors through a more direct and statistically driven lens. Within this, they consistently represent the demographic of minors as perpetrators, framing them directly through the lens of their crimes and consequences, detailing specific charges

and disciplinary actions; further substantiated with national statistics, underscoring that a significant proportion of perpetrators in these deepfake crimes are minors. This explicit and statistical framing of minor perpetration establishes them as clear agents of harm, making their involvement a central, undeniable aspect of the public discourse, reflecting a nation already deeply engaged in legislating and institutionalising responses to this specific form of digital violence, implicitly acknowledging the broader systemic issues that underpin such crimes, allowing for a societal discourse that supports comprehensive, proactive regulatory and institutional responses.

However, even with this more nuanced approach, South Korean media's discourse still highlights ongoing challenges in addressing deepfake harms holistically. While critical of the execution of existing processes, such as the "secondary victimization" from delayed institutional responses, their coverage notably does not convey an absence of legal or institutional frameworks. Instead, a shortcoming evident in South Korean representation is its near silence on the technology companies' systemic role in deepfake creation or prevention in their media coverage. While they *are* critiqued for failing to promptly remove harmful content despite having policies, The Independent (UK - Source 7) was the only article to vociferously call for mandated action, using the victim's struggle to get content removed from Snapchat as a central driver for proposed legislation. In other articles, despite acknowledging the ease of distribution on social media, platform responsibility for regulating distribution or implementing preventative measures remains largely implicit. This constitutes a blind spot regarding corporate accountability, even within proactive regulatory environments like South Korea, pointing to a persistent challenge in fully addressing the architecture of exploitation where the very tools of harm are developed. It suggests a societal reluctance to apply the full force of accountability to the creators and facilitators of the informatics of domination at their source, thereby limiting the comprehensiveness of global regulatory efforts (Haraway, 2016, p.29-33).

Finally, the volume and emphasis of media representation itself further reveals a key component of the required response: sustained public and political will. Searching for relevant articles revealed a notably limited and inconsistent volume of media representation regarding "deepfake" and "minors" across international (Al Jazeera, Deutsche Welle, and Reuters) and Western news outlets. In the US, several major outlets returned no relevant articles, and even CNN searches required navigating unrelated content to find minor-related deepfake material. This demonstrated a difficulty in finding substantial representation of minors' dual role, contradicting the severity and urgency occasionally framed within the articles themselves. The UK exhibited similar patterns, with the BBC only explicitly covering the dual role in coverage of the 2024 South Korean case, despite domestic recognition of the issue. This scarcity in Western media aligns with the digital immigrant tendency to either misinterpret or downplay the full scope of emerging digital harms that do not fit traditional categorizations, thus failing to generate the necessary public and political momentum for comprehensive responses (Prensky, 2001, p.2).

In sharp contrast, South Korean media provided significantly more representation, even when searched in English. Further investigation using translated search terms[8] on Naver - South Korea's equivalent of Google and a major news aggregator - revealed a constant stream of reporting. Searching terms like 'Deepfake Technology' [딥페이크 기술] immediately presented numerous articles, many published within recent weeks or days, centrally focused on framing the minor perpetration crisis, often highlighted directly in titles. This pervasive coverage, spanning every type of outlet - from financial and political to social, mainstream news, and even blogs and op-eds - demonstrates the deepfake issue is a deeply central aspect of current South Korean public discourse, accurately reflecting the national statistics of significant minor perpetration.

The contrast between the high volume of explicit, perpetrator-focused reporting in South Korean media versus limited, cautious coverage in Western [UK/US] media appears to correlate directly with the comprehensiveness and urgency of national legislative responses (Idealseo Reporter, 2025, para. 13-21; Seok et al., 2025, p.2-9). South Korea's more proactive and multi-dimensional measures align with its constant media stream, which keeps the issue front and centre, suggesting that consistent, explicit media framing are not merely a reflection of a problem, but a catalyst for the public and political momentum necessary to implement complex, proactive, and effective solutions. Effective counter-discourses can challenge the architecture of exploitation and actively influence the regulatory landscape within the informatics of domination, particularly those that address the nuanced roles of minors and the foundational harms of deepfake creation.

[8] Child – 어린이; Children – 어린이들; Minors – 미성년자; Deepfake Technology - 딥페이크 기술; School – 학원 // 학교

6. The Digital Violation Discrepancy - Centring Marginalised Experiences to Address Deepfake Harms

The following chapter interprets the empirical findings from the focus groups and CDA through the previously established virtual feminist theoretical framework, directly addressing the primary research question: *How are the impacts and perceptions of deepfake technologies related to the underrepresentation of women in AI development and regulation?* Findings reveal a Digital Violation Discrepancy, implying that the unique and disproportionate harms women endure are not mere coincidences, but systemic consequences of a technological landscape fundamentally shaped by the underrepresentation of female voices in AI development and regulation. Further in-depth analysis and a comprehensive exploration of the implications are presented in *Appendix C* to accommodate word count stipulations.

The core of this discrepancy lies in differing understandings of deepfake harms, influenced significantly by gender and generational perspectives. Female participants consistently identified the non-consensual *creation* of deepfakes as the initial cause of harm, perceiving it as an irrefutable violation of personal autonomy and bodily integrity, regardless of subsequent dissemination. This perspective, deeply rooted in the lived experiences of women and their historical marginalisation from technology's inception, underscores how deepfakes impose a fabricated digital performance onto an individual, directly assaulting their agency. In contrast, male discourse prioritised *dissemination* as the primary source of harm, focusing on visible, external, or legal consequences; this male-centric understanding, reflective of a predominantly male creator base in AI, contributes to an oversight of the initial, intimate violation for women, leaving these fundamental harms largely unaddressed in public discourse and subsequent policy. For a comprehensive discussion of these findings and their interpretation through a virtual feminist lens, see *Appendix C. II*.

This divergence in perceiving harm has significant implications for regulatory responses, often leading to a miscalibration in their design. Current legislative efforts largely maintain a traditional punitive focus on punishment for dissemination rather than prevention at the source. However, deepfakes introduce unique challenges to this paradigm, particularly around the anonymity afforded to perpetrators through privacy-forward networks and the availability of services designed to erase digital footprints. This circumvention of traditional judicial oversight highlights that punishment alone is an insufficient deterrent when the act can be easily hidden or untraceable. A detailed analysis of these regulatory challenges and

their differential framing across regions, including insights from South Korea, is provided in Appendix *C.II*.

Furthermore, the research points to an unaddressed ethical void concerning platform complicity. Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) are frequently implicated in facilitating AIG-NCII creation and distribution, yet often operate without sufficient accountability or ethical safeguards designed to mitigate harms disproportionately affecting marginalised groups. The continued allowance of user-friendly deepfake tools and marketplaces, despite a statistical recognition of the rise of AIFA crimes and calls for prevention from female participants, suggests a societal negligence towards pervasive public access that cannot be excused by the impossibility of absolute private containment. A full discussion on platform responsibility, regulatory failures, and the gendered implications of these ethical voids is presented in *Appendix C. III*.

Finally, the findings highlight a critical paradox concerning minors: their significant involvement as both vulnerable victims and perpetrators. Societal ambiguities around minor sexualisation, reflected in arguments defending AIG-CSAM as ‘harmless fantasy’, perpetuate a dangerous normalisation of digital violence. While Western media narratives often hesitate to frame minor deepfake involvement as a crime, contrasting with South Korea's more stringent approach, the ongoing challenge of anonymity suggests that a holistic approach, combining punitive measures with fundamental education on consent and digital ethics, is crucial. For an in-depth examination of the complex realities of minor involvement and the ethical challenges they pose, refer to *Appendix C. IV*.

In summary, the impacts and perceptions of deepfake technologies are inextricably linked to the underrepresentation of women in AI development and regulation. This systemic bias shapes how harms are understood, addressed, and legislated, perpetuating a Digital Violation Discrepancy that disproportionately affects women and minors. Addressing these violations necessitates centring marginalised experiences as a requirement in AI development and regulation to build systems that reflect a more inclusive and ethical understanding of digital harm.

7. Conclusion: Toward an Equitable and Victim-Centred Digital Future

In a world captivated by AI advancements, it's easy to get caught up in the excitement and overlook the significant harms these technologies can generate. This thesis closely examined at one such challenging area: AI-generated deepfakes. The research focused on their widespread impact, the disproportionate harm they inflict on women and girls, and the dual role of minors amidst this. Combining focus groups insights and a critical analysis of media, grounded in virtual feminist theories, the research has uncovered complex ways in which digital violence plays out in our fluctuating technological world. Findings consistently show that deepfakes are not just technical curiosities; they are powerful tools of technological sexual violence, highlighting clear gender differences in how harm is perceived and revealing the complicated roles young people play in this urgent global issue.

Through the emergent Digital Violation Discrepancy theory, this research directly answers the original research question: The underrepresentation of diverse voices, especially women, in AI development and regulation does shape the impacts and perceptions of deepfake technologies, leading to unique and disproportionate harms for women. As explored in Chapter 6 and comprehensively analysed with supporting empirical evidence and theoretical application in Appendix B, this problem stems from a technological landscape imbued with patriarchal biases, causing a fundamental to disconnect in understanding deepfake harms. This deep-rooted discrepancy manifests in several critical ways. Firstly, the fundamental consent violation that occurs at the point of creation for women is often overlooked or downplayed by a male-centric discourse, which instead prioritises visible harms related to distribution. Secondly, this underrepresentation creates an ethical void in platform accountability; large online platforms, often built without diverse ethical compasses, fail to proactively prevent harm or face sufficient external pressure, allowing an architecture of exploitation to flourish and inflict unique harms on women. Lastly, findings highlight a widespread misunderstanding of digital native behaviours and the dangerous normalisation of fabricated harm, like AIG-CSAM. Western media and regulatory bodies, operating from a digital immigrant perspective, struggles to grasp the nuances of minors' dual roles as both victims and perpetrators, leading to superficial analyses that miss underlying societal factors, such as male resistance during periods of gender reordering. Ultimately, addressing this Digital Violation Discrepancy requires a fundamental shift: centring women's experiences in all stages of AI development and regulation is not just about fairness, but is essential for truly understanding and combating

deepfake harms to build an equitable digital future for everyone. For a comprehensive outline of policy recommendations derived from this research, please refer to Appendix D.

7.1 Limitations

Despite garnering insightful findings, this research has several limitations that should be acknowledged. The qualitative nature of focus group research, with a relatively small sample of 14 participants, naturally limits the findings' broad applicability. While the data offered rich, detailed insights, its small scale makes generalisability more difficult. Furthermore, as aforementioned, the predominantly white-European composition of the focus group participants suggests the findings primarily reflect a specific cultural perspective; this homogeneity limited thorough intersectional analysis, potentially overlooking how other identity markers (such as race, socio-economic background, or LGBTQ+ status) might combine with gender to influence perceptions and experiences of deepfake harms. As such it offers an opportunity for future research to aim for greater cultural diversity, capturing a wider range of experiences. Additionally, although a non-directive approach was used, the researcher personally facilitating the all-male focus group, contrary to initial plans for an outside facilitator, could have theoretically introduced subtle influences on group dynamics or participants' candidness, though observations during the session suggested high engagement. Finally, the CDA focused on a limited selection of seven mainstream news articles from specific geographic regions. While these were chosen for relevance and to represent public discourse, this selection might not cover the full variety of media narratives, including those from niche outlets or non-Western regions with different cultural contexts and media reporting styles.

7.2 Avenues for Future Research

The thesis's insights and limitations open several promising avenues for future research to deepen understandings of AI-generated deepfakes and help develop more effective solutions. As mentioned, future studies should consider broader generational and intersectional analyses, replicating the focus groups with more diverse participants, including older demographics to bridge the digital immigrant gap, and racially, socio-economically, and culturally varied groups to provide deeper insights into how diverse lived experiences shape understandings of deepfake harms. It would also be highly valuable, though ethically sensitive, to conduct in-depth studies with perpetrators, particularly former minor perpetrators of deepfake crimes, to gain unprecedented insights into their motivations, reasoning, perception of consequences, and the role of online subcultures in normalising

such behaviour - essential for developing effective prevention and rehabilitation strategies. Comparative policy and regulatory analyses across different nations (e.g., South Korea, China, EU member states, and countries in the Global South) could identify best practices, assessing the efficacy of various legal and technological approaches, and exploring challenges in harmonising international responses. Further, research into technological interventions and ethical AI development is encouraged, focusing on developing and evaluating the effectiveness of technical safeguards, like robust detection algorithms, proactive blocking systems, and verifiable consent frameworks within AI models and platforms designed to prevent non-consensual deepfake creation, potentially involving collaborations among ethicists, legal experts, and AI engineers. Evaluation of current and proposed educational interventions should involve developing, implementing, and rigorously assessing digital literacy and consent education programs in schools and communities to gauge their effectiveness in raising awareness, promoting ethical online behaviour, and mitigating the normalisation of digital harm among young people. Finally, further research into the design and implementation of effective platform accountability mechanisms is crucial, including independent oversight bodies, transparency requirements, and legal frameworks that compel proactive moderation and swift removal of harmful content across platforms, while also exploring the challenges and potential solutions for securing cooperation from privacy-focused messaging applications. By pursuing these areas, future research can build on the foundational understanding established in this thesis, contributing to a more nuanced, evidence-based, and ultimately, a fairer and safer digital landscape for everyone.

8. Reference List

Aavik, K., Collinson, D. L., Hall, M., & Hearn, J. (2024, February). *The Impact of Men's Domination of AI and Deepfake Technology*. Research Gate. https://www.researchgate.net/publication/379085442_The_Impact_of_Men's_Domination_of_AI_and_Deepfake_Technology_Violations_of_Women_Girls_and_Democracy_in_a_Digital_Age

the original resource has been removed: https://www.routledge.com/blog/article/open-ai-another-case-of-men-masculinities-gendered-organizingbut it the pdf is still online

Ajder, H., Patrini, G., Cavalli, F., Cullen, L., & Deeptrace. (2019). The State of Deepfakes: landscape, threats, and impact. In *Deeptrace* [Report]. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Akselrod, O., & Venzke, C. (2025, February 11). Trump's efforts to dismantle AI protections, explained | ACLU. *American Civil Liberties Union*. https://www.aclu.org/news/privacy-technology/trumps-efforts-to-dismantle-ai-protections-explained

Amidi, A. (2024, January 31). *New report confirms worst fears: AI will disrupt countless animation jobs over next 3 years*. Cartoon Brew. https://www.cartoonbrew.com/artist-rights/union-study-says-generative-ai-will-disrupt-204000-jobs-three-years-237495.html

Atherton, D. (2024, September 1). *Deepfakes and Child Safety: A survey and analysis of 2023 incidents and responses*. AI Incident Database. https://incidentdatabase.ai/blog/deepfakes-and-child-safety/

Bang, J., & Go, K. (2024, August 28). *'n번방 대학동문 성범죄' 피해 60여명. . .경찰이 손놓자 직접 나섰다*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1141455.html

Barrabi, T. (2025, May 28). AI could spark bloodbath for white collar jobs — and send unemployment to 20%: Anthropic CEO. *New York Post*. https://nypost.com/2025/05/28/business/ai-could-cause-bloodbath-for-white-collar-jobs-spike-unemployment-to-20-anthropic-ceo/

Beaumont, R. (2022, March 31). *LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS | LAION*. LAION.ai. https://laion.ai/blog/laion-5b/

Brooks, T., Verizon, Princess G., Heatley, J., JP Morgan Chase & Co., Jeremy J., United States Secret Service, Scott Kim, Experian, Samantha M., Federal Bureau of Investigation, Sara Parks, National

Cyber-Forensics & Training Alliance, Maureen Reardon, Melian LLC, Harley Rohrbacher, Burak Sahin, Deloitte & Touche, Shani S., . . . Richard V. (2021). Increasing Threat of Deepfake Identities. In *Homeland Security*. Homeland Security. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

Brown, S. (2021, April 21). *Machine learning, explained*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Buffet Brief. (2023). The rise of artificial intelligence and deepfakes. In *BUFFET T BRIEF* (pp. 1–2) [Journal-article]. https://buffett.northwestern.edu/documents/buffett-brief_the-rise-of-ai-and-deepfake-technology.pdf

Butler, J. (1999). *Gender trouble: Feminism and the Subversion of Identity*. Theatre Arts Books.

Cambridge Dictionary. (2025). NSFW. In *English Meaning - Cambridge Dictionary*. https://dictionary.cambridge.org/dictionary/english/nsfw

Charlwood, R. (2025, February 12). *IProov study reveals deepfake blindspot: Only 0.1% of people can accurately detect AI-Generated Deepfakes* |. iProov. https://www.iproov.com/press/study-reveals-deepfake-blindspot-detect-ai-generated-content#:~:text=Key%20Findings%3A,deepfakes%20is%20likely%20even%20higher.

Choi, J.-Y. (2025, April 11). Digital sex crime victims surpass 10,000 in Korea, majority in teens, 20s - The Korea Herald. *The Korea Herald*. https://www.koreaherald.com/article/10463398

Choi, S., Lee, W., Choi, Y., Park, G., & Choi, Y. (2024, September 2). *"혹시 내 사진도?". . .학교 덮친 딥페이크 범죄 공포*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1155647.html

Choi, Y. (2024a, June 4). *유엔 "한국, 디지털성폭력 폭발적 증가. . .가해자 기소율 낮아."* 한겨레. https://www.hani.co.kr/arti/society/women/1143393.html

Choi, Y. (2024b, September 2). *'n번방 폭로' 박지현 "딥페이크, 국가 비상사태 선포해야."* 한겨레. https://www.hani.co.kr/arti/society/women/1155359.html

Choi, Y. (2024c, September 2). *서지현 "디지털성범죄 지옥문 2년 전 경고. . .국가는 뭐 했나."* 한겨레. https://www.hani.co.kr/arti/society/women/1155484.html

Choi, Y. (2024d, September 2). *집단 성범죄 통로 딥페이크, 1020엔 이미 '보통의 장난.'* 한겨레. https://www.hani.co.kr/arti/society/women/1154477.html

Choi, Y. (2024e, September 2). *"텔레그램은 못 잡아" 경찰이 한다는 말. . .피해자가 수사 나섰다.* 한겨레. https://www.hani.co.kr/arti/society/society_general/1143702.html

Choi, Y. (2024f, September 2). *"한국 여성은 나라가 없다". . .22만명 연루 딥페이크 성범죄 파문.* 한겨레. https://www.hani.co.kr/arti/society/women/1155539.html

Choi, Y., & Park, H. (2024, September 2). *"유포 목적 없다". . .만들어도 시청해도 처벌 피하는 딥페이크.* 한겨레. https://www.hani.co.kr/arti/society/women/1154765.html

Chung-Hee, S. S. (1993). Sexual equality, male superiority, and Korean women in politics: Changing gender relations in a ?patriarchal democracy? *Sex Roles*, *28*(1–2), 73–90. https://doi.org/10.1007/bf00289748

Crofts, P. (2024). Reconceptualising the crimes of Big Tech. *Griffith Law Review*, 1–25. https://doi.org/10.1080/10383441.2024.2397319

Cruz, B. (2024, September 26). *2024 DeepFakes Guide and Statistics*. Security.org. https://www.security.org/resources/deepfake-statistics/#statistics

Custers, B., & Fosch-Villaronga, E. (2022). Law and Artificial Intelligence. In *Information technology and law series/Information technology & law series*. https://doi.org/10.1007/978-94-6265-523-2

Demony, C. (2025, January 7). Britain to make sexually explicit "deepfakes" a crime. *Reuters*. https://www.reuters.com/world/uk/britain-make-sexually-explicit-deepfakes-crime-2025-01-07/#:~:text=Data%20from%20UK%2Dbased%20Revenge,creating%20and%20sharing%20these%20images.

Digital Watch. (2025, April 11). Victims of AI-driven sex crimes in Korea continue to grow | Digital Watch Observatory. *Geneva Internet Platform Digital Watch Observatory*. https://dig.watch/updates/victims-of-ai-driven-sex-crimes-in-korea-continue-to-grow

Ding, M. L., & Suresh, H. (2025). *The malicious technical ecosystem: exposing limitations in technical governance of AI-Generated Non-Consensual intimate images of adults*. Cornell University. https://arxiv.org/html/2504.17663

ECPAT International IN Online Exploitation. (2020, December 15). *Social media messaging apps host underground child sexual abuse networks - ECPAT*. ECPAT. https://ecpat.org/story/social-media-messaging-apps-host-underground-child-sexual-abuse-networks/

Elkind, D. (1967). Egocentrism in adolescence. *Child Development*, *38*(4), 1025. https://doi.org/10.2307/1127100

European Comission. (2024, August 28). *Deepfake - a global crisis*. Intellectual Property Helpdesk. https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/deepfake-global-crisis-2024-08-28_en

European Commission. (2022, October 27). *The EU's Digital Services Act*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Farkas, J., & Maloney, M. (2024). *Digital media metaphors: A Critical Introduction*. Taylor & Francis.

Flynn, A., Clough, J., & Cooke, T. (2021). Disrupting and Preventing Deepfake abuse: Exploring Criminal Law Responses to AI-Facilitated Abuse. In *Springer eBooks* (pp. 583–603). https://doi.org/10.1007/978-3-030-83734-1_29

Foucault, M. (1995). *Discipline and punish: The Birth of the Prison*. Vintage.

Galič, M., Timan, T., & Koops, B. (2016). Bentham, Deleuze and Beyond: An Overview of Surveillance Theories from the Panopticon to Participation. *Philosophy & Technology*, *30*(1), 9–37. https://doi.org/10.1007/s13347-016-0219-1

Garg, R., & Sengupta, S. (2019). "When you can do it, why can't I?": Racial and Socioeconomic Differences in Family Technology Use and Non-Use. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–22. https://doi.org/10.1145/3359165

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for Qualitative Research*. Transaction Publishers.

Go, N. (2024, September 2). *[단독] '○○○ 능욕방' 딥페이크, 겹지인 노렸다. . .지역별·대학별·미성년까지*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1154763.html#cb

Gordon, C. (2023, February 2). *ChatGPT is the fastest growing app in the history of web applications*. Forbes. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/

Grewal, D., Guha, A., & Becker, M. (2024). AI is Changing the World: Achieving the Promise, Minimizing the Peril. *Journal of Macromarketing*, *44*(4), 936–947. https://doi.org/10.1177/02761467241289573

Guajardo, A., & Tadros, E. (2024). Sexual Assault: The burden of proof for survivors. *Psychology of Woman Journal*, *5*(4), 24–36. https://doi.org/10.61838/kman.pwj.5.4.4

Gulsen, O., & Van Der Plas, C. (2025, May 6). Dutch victims overjoyed by take down of deepfake porn site. *NL Times*. https://nltimes.nl/2025/05/06/dutch-victims-overjoyed-take-deepfake-porn-site

Haggerty, K. D., & Ericson, R. V. (2003). The surveillant assemblage. *British Journal of Sociology*, *51*(4), 605–622. https://doi.org/10.1080/00071310020015280

Han, C., Li, A., Kumar, D., & Durumeric, Z. (2024). Characterizing the MrDeepFakes sexual deepfake marketplace. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.11100

Hankyoreh. (2024). *[사설] 만연한 딥페이크 성범죄, 당국이 심각성 인지해야*. 한겨레. https://www.hani.co.kr/arti/opinion/editorial/1154781.html

Haraway, D. J. (2016). A Cyborg manifesto. In *Manifestly Haraway*. University of Minnesota Press. https://warwick.ac.uk/fac/arts/english/currentstudents/undergraduate/modules/fictionnownarrativemediaandtheoryinthe21stcentury/manifestly_haraway_----_a_cyborg_manifesto_science_technology_and_socialist-feminism_in_the_. . ..pdf

Harding, S. G. (2016). *Whose science? whose knowledge?: Thinking from Women's Lives*. Cornell University Press.

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Tech Review*, *17*, 99–127. https://scholarship.law.duke.edu/dltr/vol17/iss1/4

Healey, J. (2024, August 29). Many students know of peers who created deepfake nudes, report says - Los Angeles Times. *Los Angeles Times*. https://www.latimes.com/california/story/2024-08-29/11-percent-of-students-say-classmates-have-created-deepfake-ai-nudes-report-says

Heikkilä, M. (2024, March 6). Nobody knows how AI works. *MIT Technology Review*. https://www.technologyreview.com/2024/03/05/1089449/nobody-knows-how-ai-works/

Hemrajani, A. (2023, March 8). China's New Legislation on Deepfakes: Should the Rest of Asia Follow Suit? *The Diplomat*. https://thediplomat.com/2023/03/chinas-new-legislation-on-deepfakes-should-the-rest-of-asia-follow-suit/

Hendrix, J., & Lima-Strong, C. (2025, May 22). *US House passes 10-Year moratorium on state AI laws*. Tech Policy Press. https://www.techpolicy.press/us-house-passes-10year-moratorium-on-state-ai-laws/

Higgins, R., Plunkett, C., Katz, G., Koltai, K., & De Tolly, K. (2025, January 28). *Faking It: Deepfake porn site's link to tech companies*. Bellingcat. https://www.bellingcat.com/news/uk-and-europe/2025/01/28/deepfake-porn-sites-link-to-tech-companies/

Hlavka, H. R. (2016). Speaking of stigma and the silence of shame. *Men And Masculinities*, *20*(4), 482–505. https://doi.org/10.1177/1097184x16652656

Hobbes, T. (2009). *Leviathan*. https://books.googleusercontent.com/books/content?req=AKW5Qacb7Ke-p5_ReegoHgB__WqiFuieNYJMI6tZe7lH0VhtJNoKa6xqm67XaKcBBGS8OlBW4GFkDz1pAL74NdIMgEwV36uGbFlV4TQc-nNSP94qWLuX9_MB8ghPIKWExLKroEzcedKGfce3lRredxb7W17CNkVX3orTZt0LSmHU2L5V5EGid3vYxe92vGSEfSXkFCEPsf8zWRD921L6MOKgjq8zoYItJx4-HirQVmDzFo1POzZ8dCWVy4-2L9ulPx1VUXTuF4P0mQKR_k-0mnWFlTdQVEaUCw

Hörnle, J. (2025, January 23). *Deepfakes and the Law: Why Britain needs Stronger Protections against Technology-Facilitated Abuse*. Queen Mary University of London. https://www.qmul.ac.uk/media/news/2025/humanities-and-social-sciences/hss/deepfakes-and-the-law-why-britain-needs-stronger-protections-against-technology-facilitated-abuse.html

Howard, K. (2020, October 20). *Deconstructing Deepfakes—How do they work and what are the risks?* U.S. Government Accountability Office. https://www.gao.gov/blog/deconstructing-deepfakes-how-do-they-work-and-what-are-risks

Hsu, T. (2023, January 22). As Deepfakes Flourish, Countries Struggle With Response. *New York Tikes*. https://www.nytimes.com/2023/01/22/business/media/deepfake-regulation-difficulty.html

Idealseo Reporter. (2025, May 29). 청소년 불법약물 차단 강화. . .딥페이크 성범죄물 24시간내 삭제. *연합뉴스*. https://www.yna.co.kr/view/AKR20250529001200530?input=1195m

Im, C. (2024, August 27). 서거석 전북교육감 "딥페이크 등 사이버범죄 대응 강화하라." *뉴스1*. https://www.news1.kr/local/jeonbuk/5522937

Im, I. (2024, September 1). "잡힐 리 없어", "안심하라". . .단속 비웃는 딥페이크 가해자들. *연합뉴스*. https://www.yna.co.kr/view/AKR20240830085700004

Internet Watch Foundation. (2024). [AI CSAM REPORT UPDATE]. In *AI CSAM REPORT UPDATE* [Report]. https://admin.iwf.org.uk/media/opkpmx5q/iwf-ai-csam-report_update-public-jul24v11.pdf

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information Communication & Society*, *27*(6), 1095–1109. https://doi.org/10.1080/1369118x.2023.2234980

Jeong, I. (2024, September 2). *"학교 밖 가해자" "입결 악영향". . .대학들, 딥페이크 '피해자 탓.'* 한겨레. https://www.hani.co.kr/arti/society/women/1155075.html

Jo, J., Kim, L., Noh, J. in K.-H. 2024, & In, K.-H. (2024). *"딥페이크 성범죄 피의자 73%가 10대". . . 뒤늦은 대책 마련*. Naver. https://n.news.naver.com/article/021/0002656614?sid=102

Jo, Y. (2024, September 2). *'n번방 대학동문 성범죄' 피해자 "주변 의심하게 돼 더 공포."* 한겨레. https://www.hani.co.kr/arti/society/society_general/1141734.html

JTBC. (2024, August 26). *[단독] 전국 학교 230곳에 '딥페이크방'. . ."얼굴 나온 거 다 지워라" 패닉*. Nate News. https://news.nate.com/view/20240826n32713

Jung, H. W. (2023). A new variation of modern prejudice: young Korean men's anti-feminism and male-victim ideology. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1230577

Jung, J.-H. (2025, April 10). Deepfake, AI-related digital sex crimes targeting women, children surge in Korea. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/news/2025-04-10/national/socialAffairs/Deepfake-AIrelated-digital-sex-crimes-targeting-women-children-surge-in-Korea/2282432

Kamal, A. J. (2024, November 6). *A Culture of Shame and Regret: Exploring the rise of digital sex crimes in South Korea – UAB Institute for Human Rights Blog*. https://sites.uab.edu/humanrights/2024/11/06/a-culture-of-shame-and-regret-exploring-the-rise-of-digital-sex-crimes-in-south-korea/

KeyNorth Group. (2023, May 17). *The Clearnet, the deep web and the dark Web - KeyNorth Group - Open source intelligence training*. KeyNorth Group - Open Source Intelligence Training. https://osinttraining.net/guide/osint-backgrounder/the-clearnet-the-deep-web-and-the-dark-web/

Killean, R., McAlinden, A., & Dowds, E. (2022). Sexual violence in the digital Age: Replicating and augmenting harm, victimhood and blame. *Social & Legal Studies*, *31*(6), 871–892. https://doi.org/10.1177/09646639221086592

Kim, C. (2024, September 19). 아동청소년 딥페이크 처벌 강화. . .협박 징역 3년·강요 5년 이상. *연합뉴스*. https://www.yna.co.kr/view/AKR20240919149100001

Kim, E. T. (2024, November 23). The rise of 4B in the wake of Donald Trump's reëlection. *The New Yorker*. https://www.newyorker.com/news/the-lede/the-rise-of-4b-in-the-wake-of-donald-trumps-reelection

Kim, J. (2024). Misogyny and gender conflicts in South Korea. *Communications in Humanities Research*, *44*(1), 19–28. https://doi.org/10.54254/2753-7064/44/20240118

Kitzinger, J. (1995). Qualitative Research: Introducing focus groups. *BMJ*, *311*(7000), 299–302. https://doi.org/10.1136/bmj.311.7000.299

Ko, N.-R. (2024, August 30). Mocking efforts to unmask them, Telegram sex offenders ramp up deepfakes. *Hankyoreh*. https://english.hani.co.kr/arti/english_edition/e_national/1156218.html

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Kollock, P., Blumstein, P., & Schwartz, P. (1985). Sex and Power in Interaction: Conversational Privileges and Duties. *American Sociological Review*, *50*(1), 34. https://doi.org/10.2307/2095338

Koltai, K., Sheldon, M., & Patin, N. (2025, March 25). *Behind a Secretive Global Network of Non-Consensual Deepfake Pornography - bellingcat*. Bellingcat. https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/

Koltai, K., & Zhu, M. (2024, October 15). *AI "Art" site OpenDream let users generate CSAM*. Bellingcat. https://www.bellingcat.com/news/2024/10/14/opendream-ai-image-generation-csam-vietnam/

@KORmennow. (2024, September 25). *Screenshot of a translated South Korean blog post detailing how to wipe evidence of deepfake crimes*. X. Retrieved June 26, 2025, from https://x.com/KORmennow/status/1838823610745786868

Original blog post has since been removed

Krašna, M., & Bratina, T. (2011, May 1). *The perception of digital security among digital natives*. IEEE Conference Publication. https://ieeexplore.ieee.org/document/5967248?denied=

Krueger, R. A., & Casey, M. A. (2014). *Focus groups: A Practical Guide for Applied Research*. SAGE Publications, Incorporated.

Lageson, S. E. (2016). Digital punishment's tangled web. *Contexts*, *15*(1), 22–27. https://doi.org/10.1177/1536504216628841

Lakatos, S. (2023). A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business. In *Graphika* [Report]. https://22006778.fs1.hubspotusercontent-na1.net/hubfs/22006778/graphika-report-a-revealing-picture.pdf?utm_campaign=Report%20Demand%20Gen&utm_medium=email&_hsenc=p2ANqtz-_A4Chhn2ctwpkWPuWLDBB6VeV0rx0ONYkhVm7ZPxX_J624GxzsYwYqeYInI121dUB_BFSYUXgLqkOugieuoMpoDoH6aA&_hsmi=295744348&utm_content=295744348&utm_source=hs_automation

Langdridge, D., Flowers, P., & Carney, D. (2023). Male survivors' experience of sexual assault and support: A scoping review. *Aggression and Violent Behavior*, *70*, 101838. https://doi.org/10.1016/j.avb.2023.101838

Lee, H. (2024a, August 27). More Koreans fall victim to deepfake sex crimes. *The Korea Times*. https://www.koreatimes.co.kr/southkorea/law-crime/20240826/more-koreans-fall-victim-to-deepfake-sex-crimes

Lee, H. (2024b, September 1). Deepfake map made by middle school student goes viral nationwide. *The Korea Times*. https://www.koreatimes.co.kr/southkorea/society/20240830/deepfake-map-made-by-middle-school-student-goes-viral-nationwide

Lee, J. (2024, September 24). *경찰, 내년 3월까지 '딥페이크 성범죄' 특별 집중단속*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1155567.html

Lee, J., Oh, Y., & Park, G. (2024, September 2). *'떴다방' 같은 텔레그램 성범죄, 온라인 잠입보다 강력한 대안 절실*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1154952.html

Lee, J.-H. (2024, September 2). *서울대 이어 또 대학 '딥페이크 성범죄물' 공유방. . .참가자 1200명*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1154391.html

Lee, J.-J. (2024, August 28). [Online Predators] Deepfake pornography haunts S. Korea - The Korea Herald. *The Korea Herald*. https://m.koreaherald.com/article/3463420

Lee, J.-J. (2025, March 5). Deepfake pornography crimes in South Korea shift to other encrypted messaging apps. *Asia News Network*. https://asianews.network/deepfake-pornography-crimes-in-south-korea-shift-to-other-encrypted-messaging-apps/

Lee, W. (2024, September 2). *딥페이크 피해, 학교로 확산. . .교사노조 "국가 차원 신고 받아야."* 한겨레. https://www.hani.co.kr/arti/society/schooling/1155378.html

Lee, W., Choi, S., Park, G., & Choi, Y. (2024, September 2). *"혹시 내 사진도?". . .학교 덮친 딥페이크 범죄 공포*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1155647.html

Lee, Y. (2024, August 28). *이준석 "딥페이크, 대통령 관심에 과잉규제 우려. . .불안 과장 안 돼" [영상]*. Hani. https://www.hani.co.kr/arti/politics/politics_general/1155613.html

Lim, C. (2024, August 27). *[단독]22만 "딥페이크" 텔레방에 이어 40만 유사 텔레방 확인*. Naver. https://n.news.naver.com/article/003/0012750768

Lim, J. (2024, August 26). *'디지털 피난처'로 살찌운 텔레그램, 성착취 '디지털 온상'이 되다*. 한겨레. https://www.hani.co.kr/arti/economy/marketing/1155212.html

Mahmud, B. U., & Sharmin, A. (2021). Deep Insights of Deepfake Technology : A review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2105.00192

Makken, A. (2024, May 28). 4B movement. *Internationale Vrouwendag*. https://www.internationale-vrouwendag.nl/4b-movement/

Massumi, B. (1986). *Realer than Real: The Simulacrum According to Deleuze and Guattari*. Web Archive. https://web.archive.org/web/20100523090754/http://www.anu.edu.au/hrc/first_and_last/works/realer.htm

McKinnon, A. M., Mattson, R. E., & Lofgreen, A. M. (2024). Does no mean no? Situational and dispositional factors influence emerging adult men's intentions to use assault tactics in response to women's sexual refusal during hookups. *Sexual Abuse*, *37*(1), 88–118. https://doi.org/10.1177/10790632241268527

McLelland, M. J. (2005). The world of Yaoi: the internet, censorship and the global "Boys' love" fandom. *Arts and Humanities Commons*. https://ro.uow.edu.au/artspapers/147

Meyer, C. (2025, May 9). Researchers: More Accessible Deepfake Generators Fuel Rapid Increase in AI Explicit Images. *ASIS International*. https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2025/may/deepfake-generators/#:~:text=Researchers%20from%20the%20Oxford%20Internet,were%20available%20for%20public%20download.

Miller, J. B. (1985). Patterns of Control in Same-Sex Conversations: Differences between Women and Men. *Women S Studies in Communication*, *8*(2), 62–69. https://doi.org/10.1080/07491409.1985.11089680

Montemurro, B., Bartasavich, J., & Wintermute, L. (2014). Let's (Not) Talk about Sex: The Gender of Sexual Discourse. *Sexuality & Culture*, *19*(1), 139–156. https://doi.org/10.1007/s12119-014-9250-5

Morgan, D. L. (1996). Focus groups. *Annual Review of Sociology*, *22*(1), 129–152. https://doi.org/10.1146/annurev.soc.22.1.129

Mullin, J. (2025, March 5). *The TAKE IT DOWN Act: a flawed attempt to protect victims that will*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2025/02/take-it-down-act-flawed-attempt-protect-victims-will-lead-censorship#:~:text=The%20Bill%20Will%20Lead%20To%20Overreach%20and%20Censorship&text=The%20takedown%20provision%20applies%20to,misuse%20of%20its%20takedown%20regime.

namuwiki. (2025, May 15). *Deepfakes*. Namuwiki. https://namu.wiki/w/%EB%94%A5%ED%8E%98%EC%9D%B4%ED%81%AC

Narayanan, M. (2024, March 18). *The UK's Online Safety Act is not enough to address Non-Consensual deepfake pornography*. Tech Policy Press. https://www.techpolicy.press/the-uks-online-safety-act-is-not-enough-to-address-nonconsensual-deepfake-pornography/

National Sexual Violence Resource Center [NSCRC]. (2025). *Understanding male socialization, stigma, and reactions to sexual violence*. National Sexual Violence Resource Center. https://www.nsvrc.org/working-male-survivors-sexual-violence/Understanding

NCMEC. (2025). *Generative AI CSAM is CSAM*. National Center for Missing & Exploited Children. https://www.missingkids.org/blog/2024/generative-ai-csam-is-csam

Obadia, S. (2024, May 8). *Survivor safety: Deepfakes and the negative impacts of AI technology*. Maryland Coalition Against Sexual Assault | MCASA. https://mcasa.org/newsletters/article/survivor-safety-deepfakes-and-negative-impacts-of-ai-technology

Oh, S. (2024). *중·고생 3.9%, 성적 이미지 공유 요구 받아. . .불법 촬영 피해 2.7%*. 한겨레. https://www.hani.co.kr/arti/society/women/1147687.html

Ortutay, B. (2025, May 20). *Trump signs the Take It Down Act. What is it? | AP News*. AP News. https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615

Pal, K. K., Piaget, K., Zahidi, S., & Baller, S. (2024). Global Gender Gap Report 2024. In *World Economic Forum*. https://www.weforum.org/publications/global-gender-gap-report-2024/

Pal, S., Lazzaroni, R. M., & Mendoza, P. (2024, October 10). *AI's missing link: the gender gap in the talent pool*. Interface. https://www.interface-eu.org/publications/ai-gender-gap

Papadopoulos, L. (2010). Sexualisation of Young People review. In *Sexualisation of Young People Review* (pp. 3–100) [Review]. https://core.ac.uk/download/pdf/4160055.pdf

Park, G. (2024a, September 2). *[단독] 딥페이크 텔레방에 22만명. . .입장하니 "좋아하는 여자 사진 보내라."* 한겨레. https://www.hani.co.kr/arti/society/society_general/1154764.html

Park, G. (2024b, September 4). *[단독] "내 제자가 딥페이크 범인". . .가해·방관자 뒤섞인 무참한 교실*. 한겨레. https://www.hani.co.kr/arti/society/society_general/1155781.html

Park, H. (2024a, August 28). *'서울대 n번방' 딥페이크 제작 공범, 징역 5년. . ."굴욕적이고 역겨워."* 조선일보.

https://www.chosun.com/national/court_law/2024/08/28/QSV77BH47FEOTMKU6SEOCI46OI/

Park, K. A. (1993). Women and Development: The case of South Korea. *Comparative Politics*, *25*(2), 127. https://doi.org/10.2307/422348

Park, S. (2024b, January 29). *"이번 시즌 먹잇감인가요?" 놈들이 텔레그램에서 웃었다 | 셜록*. 셜록.

https://www.neosherlock.com/archives/25004

Park, S. (2024c, February 2). *피해자가 잡은 "서울대 딥페이크" 용의자, 경찰이 풀어줬다 | 셜록*. 셜록.

https://www.neosherlock.com/archives/25298

Park, S., Sung, Y., & Kim, M. (2024, November 7). *S. Korean gov't unveils stricter measures to combat deepfake sex crime*. The Chosun Daily. https://www.chosun.com/english/national-en/2024/11/07/5HRDOV75LVHKDFVDNMUNMS6ZVM/

Petsenidou, S. (2024, August 30). *The dark Underbelly of K-Pop: unraveling the "Nth Room" and its new horrors*. 저널인뉴스 - Jinkorea. https://jinkorea.kr/news/view.php?no=5792

Plant, S. (1997). *Zeros + ones: Digital Women + the New Technoculture*. Doubleday Books. https://monoskop.org/images/1/17/Plant_Sadie_Zeros_and_Ones_Digital_Women_and_the_New_Technoculture_1997.pdf

Prensky, M. (2001). Digital Natives, Digital Immigrants Part 1. *On The Horizon the International Journal of Learning Futures*, *9*(5), 1–6. https://doi.org/10.1108/10748120110424816

Ramos, G. (2022, August 20). Why we must act now to close the gender gap in AI. *World Economic Forum*. https://www.weforum.org/stories/2022/08/why-we-must-act-now-to-close-the-gender-gap-in-ai/

Randall, M. (2010). Sexual Assault Law, Credibility, and "Ideal Victims": consent, resistance, and victim blaming. *Canadian Journal of Women and the Law/Revue Femmes Et Droit*, *22*(2), 397–433. https://doi.org/10.3138/cjwl.22.2.397

Reuters. (2024, August 27). *South Korean President Yoon proposes a new dialogue channel with North Korea* [Video]. NBC News. https://www.nbcnews.com/news/world/south-korea-vows-tougher-stance-outcry-sexual-deepfakes-telegram-rcna168361#

Salvaggio, E. (2024, January 2). *LAION-5B, stable diffusion 1.5, and the original sin of generative AI*. Tech Policy Press. https://www.techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai/

Saner, E. (2024, January 31). Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box.' *The Guardian*. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

Savino, M. (2024, January 31). Here's what is being done to prevent so-called deep fakes in Connecticut. *NBC Connecticut*. https://www.nbcconnecticut.com/news/politics/heres-what-is-being-done-to-prevent-so-called-deep-fakes-in-connecticut/3207841/

Sayer, J. (2016, March 10). What the paps did to Emma Watson on her 18th birthday is so gross. *Cosmopolitan*. https://www.cosmopolitan.com/uk/entertainment/news/a41853/what-the-paps-did-to-emma-watson-on-her-18th-birthday-is-so-gross/

Seipp, T. J. (2023). Media Concentration Law: Gaps and Promises in the Digital Age. *Media and Communication*, *11*(2), 392–405. https://doi.org/10.17645/mac.v11i2.6393

Seok, H., Oh, S., Kim, S. E., & Tae, S. (2025). The Current Status and the Implications of South Korea's Response to Deepfake Sexual Crimes. In *ECPAT Korean Tacteen Naeil*. https://www.tacteen.net/attach_file/?n=31819

Shi, Y., Kiley, K., & DiPietro, S. M. (2024). To the Extreme: Exploring the Rise of a Deviant Culture in a Misogynist Digital Community. *Socius Sociological Research for a Dynamic World*, *10*. https://doi.org/10.1177/23780231241272681

Singer, N. (2024a, April 8). Teen Girls Confront an Epidemic of Deepfake Nudes in Schools. *New York Times*. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html

Singer, N. (2024b, April 20). Spurred by Teen Girls, States Move to Ban Deepfake Nudes. *New York Times*. https://www.nytimes.com/2024/04/22/technology/deepfake-ai-nudes-high-school-laws.html

Singh, S. (2024, December 30). 18 million visitors from this country use AI websites to unclothe people virtually. *Hindustan Times*. https://www.hindustantimes.com/trending/18-million-visitors-from-this-country-alone-use-ai-sites-to-unclothe-people-virtually-101735545459450.html

Somers, M. (2020, July 21). *Deepfakes, explained*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Takeuchi, C. L. (2015). REGULATING LOLICON: TOWARD JAPANESE COMPLIANCE WITH ITS INTERNATIONAL LEGAL OBLIGATIONS TO BAN VIRTUAL CHILD PORNOGRAPHY. In *Georgia Journal of International and Comparative Law* (Vol. 44, pp. 197–236). https://georgia-international-journal.scholasticahq.com/api/v1/articles/3477-regulating-lolicon-toward-japanese-compliance-with-its-international-legal-obligations-to-ban-virtual-child-pornography.pdf

Tenor, C., & Himma-Kadakas, M. (2023). Voiceless Youth – Reasons (Not) to involve Minors in news coverage. *Journalism Practice*, *18*(1), 137–157. https://doi.org/10.1080/17512786.2023.2206810

Thiel, D. (2023). Identifying and eliminating CSAM in generative ML training data and models. *Stanford Digital Repository*. https://doi.org/10.25740/kh752sm9123

Thiel, D., Stroebel, M., & Portnoff, R. (2023, June 24). *Generative ML and CSAM: Implications and mitigations*. FSI. https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations

Thorn. (2025a). Deepfake Nudes & Young People Navigating a new frontier in technology-facilitated nonconsensual sexual abuse and exploitation. In *Thorn*. https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf

Thorn. (2025b, March 14). *Mitigating the risk of generative AI models creating Child Sexual Abuse Materials — An analysis by child safety nonprofit Thorn - Partnership on AI*. Partnership on AI. https://partnershiponai.org/thorn-framework-case-study/

Trombevski, S. (2025, January 27). *Deepfakes and the Future of AI Legislation: Overcoming the Ethical and Legal Challenges*. GDPR Local. https://gdprlocal.com/deepfakes-and-the-future-of-ai-legislation-overcoming-the-ethical-and-legal-challenges/

u/Inner_Response_1714. (2024, October). *Reddit post in r/IncelTears displaying a screenshot of a translated blog post by a UOS student*. Reddit. Retrieved June 26, 2025, from https://www.reddit.com/r/IncelTears/s/lB02YnK5B9

Original blog post has since been removed

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries [PDF]. In *Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries* (pp. 1–20). https://doi.org/10.1145/3613904.3642382

United Nations Interregional Crime and Justice Research Institute [UNICRI]. (2024). Generative AI: A New Threat for Online Child Sexual Exploitation and Abuse. In *United Nations Interregional Crime and Justice Research Institute*. United Nations Interregional Crime and Justice Research Institute. https://unicri.it/sites/default/files/2024-09/Generative-AI-New-Threat-Online-Child-Abuse.pdf

Uren, T. (2025, May 23). *Telegram Is Cooperating with Authorities, for Now*. Default. https://www.lawfaremedia.org/article/telegram-is-cooperating-with-authorities--for-now

u/SouthKorea_Team_CAT. (2024, November). *Post from r/Feminism displaying a screenshot of translated South Korean News article comments*. Reddit. Retrieved June 26, 2025, from https://www.reddit.com/r/Feminism/s/JGChy36rpT

Original article:

https://www.hankookilbo.com/News/Read/A2024092114480002918?did=NA

Van Den Hoven, J., Stauch, M., Musiani, F., Domingo-Ferrer, J., Ruggieri, S., Pratesi, F., Trasarti, R., & Comandé, G. (2024). Democracy in the digital age. *shs.hal.science*. https://doi.org/10.15488/18259

Van Dijk, T. A. (1998). *Ideology ; A Multidisciplinary Approach*. Sage Publications.

Vazquez, L. (2021). RECOMMENDATIONS FOR REGULATION OF DEEPFAKES IN THE U.S.: DEEPFAKE LAWS SHOULD PROTECT EVERYONE NOT ONLY PUBLIC FIGURES. In *BIBLIOGRAPHY*. https://www.ebglaw.com/assets/htmldocuments/uploads/2021/04/Reif-Fellowship-2021-Essay-2-Recommendation-for-Deepfake-Law.pdf

Vecchietti, G., Liyanaarachchi, G., & Viglia, G. (2024). Managing deepfakes with artificial intelligence: Introducing the business privacy calculus. *Journal of Business Research*, *186*, 115010. https://doi.org/10.1016/j.jbusres.2024.115010

Williams, K. (2024a, September 13). *US states struggle to define "Deepfakes" and related terms as technically complex legislation proliferates*. Tech Policy Press. https://www.techpolicy.press/us-states-struggle-to-define-deepfakes-and-related-terms-as-technically-complex-legislation-proliferates/

Williams, K. (2024b, October 10). *Minors are on the frontlines of the sexual deepfake epidemic — here's why that's a problem*. Tech Policy Press. https://www.techpolicy.press/minors-are-on-the-frontlines-of-the-sexual-deepfake-epidemic-heres-why-thats-a-problem/

Williams, T. (2024, June 24). Deepfake abuse law 'doesn't go far enough.' *Information Age*. https://ia.acs.org.au/article/2024/deepfake-law-doesnt-go-far-enough.html

Wise, A. (2025, May 6). Major deepfake porn site shuts down. *NPR*. https://www.npr.org/2025/05/06/nx-s1-5388422/mr-deepfakes-porn-site-ai-shut-down

Wodak, R., & Meyer, M. (2015). *Methods of critical Discourse Studies*. SAGE Publications Limited.

Wolbers, H., Cubitt, T., & Cahill, M. J. (2025). Artificial intelligence and child sexual abuse: A rapid evidence assessment. In Australian Institute of Criminology, *Trends & Issues in Crime and Criminal Justice: Vol. No. 711*. https://www.aic.gov.au/sites/default/files/2025-01/ti711_artificial_intelligence_and_child_sexual_abuse.pdf

Won, E. (2024, September 6). *I saw deepfakes when exposing the Nth Room case 5 years ago — the government's lax response is to blame for their proliferation today*. Hankyoreh. https://english.hani.co.kr/arti/english_edition/e_national/1157369.html#

Wrona, A. (2025, May 14). Criminals need just 20 images of one child to produce deep fake, cyber experts say. *Irish Independent*. https://www.independent.ie/irish-news/criminals-need-just-20-images-of-one-child-to-produce-deep-fake-cyber-experts-say/a42584428.html

Yang, D. (2024, August 28). "애가 나쁜 짓". . .가해자 부모, 딥페이크 증거 지우기. *YTN*. https://www.ytn.co.kr/_ln/0103_202408281510050845_001

Yoona, C., & Park, H. (2024, September 2). *"유포 목적 없다". . .만들어도 시청해도 처벌 피하는 딥페이크*. 한겨레. https://www.hani.co.kr/arti/society/women/1154765.html

Young, J. (2021, April 3). Rapper Bhad Bhabie rakes in $1M in OnlyFans debut in under 6 hours. *Fox Business*. https://www.foxbusiness.com/lifestyle/rapper-bhad-bhabie-1-million-onlyfans-debut-six-hours

Zowghi, D., & Bano, M. (2024). AI for all: Diversity and Inclusion in AI. *AI And Ethics*, *4*(4), 873–876. https://doi.org/10.1007/s43681-024-00485-8

오세진. (n.d.). *"음란물" "성인물" 아닙니다, '딥페이크 성범죄 영상'입니다*. 한겨레. https://www.hani.co.kr/arti/society/women/1141588.html

최윤아. (n.d.). *유엔 "한국, 디지털성폭력 폭발적 증가. . .가해자 기소율 낮아."* 한겨레. https://www.hani.co.kr/arti/society/women/1143393.html

9. Appendix A: South Korea; A History of Technological Violence and the 2024 Deepfake Epidemic

This appendix provides a detailed overview of the Nth Room digital sex crime cases in South Korea, tracing their origins and evolution, with a specific focus on the widespread deepfake epidemic that emerged in 2024. While the South Korean case is not the primary focus of this thesis, a close examination of its unfolding provides important insights into social dynamics, technological vulnerabilities, and societal responses that led to the first national scale deepfake epidemic, offering valuable lessons for understanding and preventing similar occurrences in other countries.

Background: History of the Nth Rooms

The "Nth Rooms" refer to a series of highly organised and exploitative online sex trafficking rings that operated primarily on the Telegram messaging app in South Korea, first gaining significant public attention in 2019 (ECPAT, 2020, para.1-2). In these [chat]rooms, perpetrators coerced and blackmailed victims, including numerous minors, into producing sexually explicit content, which was then distributed and sold to subscribers within tiered, encrypted chatrooms, managed with varying levels of access and price (Kamal, 2024, para.8-13). The name "Nth Room" came from the successive creation of multiple chatrooms by different operators (e.g., "Room 1," "Room 2," up to "Room N"). The cases exposed the dark underbelly of digital sexual violence and prompted widespread public outrage and legislative reforms with attention to technologically mediated violence in South Korea long before the 2024 deepfake epidemic (Won, 2024, para.5-6).

As aforementioned, the original "Nth Room" case first gained major public attention in 2019, revealing a large-scale network of child and adolescent sexual exploitation crimes facilitated through Telegram chatrooms (Yoona & Park, 2-24, para. 3; Y.Choi, 2024c, para.7). Perpetrators, including individuals like "GodGod", created and distributed sexually explicit content, often involving minors who were coerced or blackmailed (Kamal, 2024, para.8-13). The public outrage spurred by these revelations led to widespread calls for stricter digital sex crime laws and robust enforcement. In response, a new punishment provision for "distribution of false video materials, etc." was added to the Special Act on the Punishment of Sexual Crimes (Sexual Violence Punishment Act) in 2020, classifying it as a digital sex crime (Kamal, 2024, para. 2, 17-22; Yoona & Park, 2-24, para. 3; Y. Choi, 2024e, para.7-8). Following the incident, the Ministry of Justice also launched a Digital Sex Crimes Expert Committee in August 2021, however, critics argue that the Digital Sex Crime TF established in the Ministry of Justice was effectively disbanded after the Yoon Seok-yeol government took office in 2022, potentially undermining preventative effort, allowing the underlying issue to persist and evolve rather than being fully curbed (Y. Choi, 2024c, para.1-2).

Despite these legislative changes and public outcry, the initial Nth Room cases highlighted challenges in law enforcement. Investigations were often suspended or not prosecuted due to

difficulties in identifying suspects - a problem exacerbated by the lack of cooperation from Telegram, whose overseas servers complicated efforts for domestic agencies to request content deletion or identify perpetrators (Bang & Go, 2024, para. 5; Y. Choi, 2024e, para. 3-7, 9-10; Go, 2024, para. 8).

The 2024 Deepfake Epidemic: Emergence and Modus Operandi

While the previous Nth Room cases primarily involved non-consensual real imagery, 2024 marked a turning point in the approaches to technologically facilitated abuse with the emergence of an epidemic of AI-generated deepfakes (J.-H. Lee, 2024, para. 2-3; JTBC, 2024, para.1). This deepfake epidemic in South Korea involved the systematic creation and distribution of sexually explicit deepfake images and videos of women and girls without their consent, once again, operating through Telegram chatrooms, offering perpetrators a degree of anonymity and impunity due to the platform's overseas servers and strong encryption (Go, 2024, para. 2-6; H. Lee, 2024b, para.1-4; H. Lee, 2024a, para. 9-12; Won, 2024, para. 24-26). The issue gained significant public attention when cases emerged from various universities, including Seoul National University and Inha University, highlighting a broader problem of digital sex crimes targeting students (JTBC, 2024, para.1; Oh, 2024b, para. 1)

The deepfake crisis began to surface prominently in May 2024, with a Telegram chatroom being discovered for sharing deepfake sexual crime materials (J. H, Lee, 2024, para. 1). This room involved 1,200 participants, primarily students from a university in Incheon, following a similar issue at Seoul National University. These chatrooms were used to create and distribute deepfake illegal composite materials by transposing women's faces onto nude photos, often accompanied by victims' personal information (including contact details and student numbers) and even deepfake audio files of them saying sexually explicit phrases (Go, 2024, para. 2-6; J. H. Lee, 2024, para. 1-5).

The first issue came to light when a victim discovered a chatroom in 2021 after being harassed online, prompting a formal complaint and police investigation (Bang & Go, 2024, para. 5-6). Despite this early warning, investigations faced repeated difficulties, with few successful arrests for distribution being made, and main operators remained unidentified, signalling an ongoing challenge in curbing the issue (J. Lee et al., 2024, para. 1-5). This forced victims to investigate their own crimes, navigating a bureaucratic and unresponsive system that frequently left them without support (Y. Choi, 2024e, para. 7-11; Y. Jo, 2024, para. 4-6; S. Park, 2024b, para. 19, 25-26).

These rooms often had organised roles for exchanging, threatening victims, and recruiting new participants (Bang & Go, 2024, para. 3-4; Won, 2024, para. 7-9). The process typically involved perpetrators gathering in Telegram channels, often referred to as "overlapping acquaintance rooms," organised by region or university (Go, 2024, para. 2-3). Within the rooms, users would then identify mutual acquaintances, share ordinary photos from social media (such as KakaoTalk profile pictures),

and manipulate these into hyper-realistic, non-consensual content (Y. Jo, 2024, para. 4; S. Park, 2024a, para. 20-22; Won, 2024, para. 16-20; Jo et al., 2024a, para 1&3). In addition to images, victims' personal information, including contact details and student numbers acting as a form of "entry fee", creating a predatory ecosystem of systematic harassment (Go, 2024, para.3-4; Won, 2024, para. 18). The sophisticated nature of these operations sometimes involved AI bots that could create deepfake material within seconds [Fig.6].

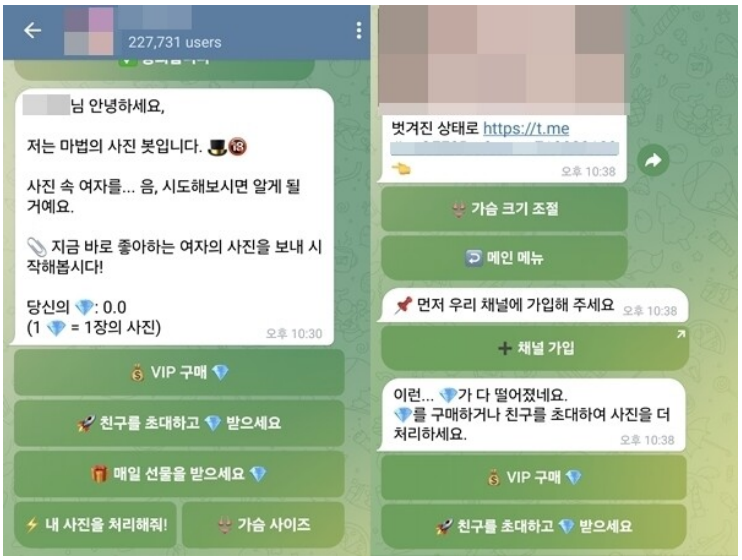

Figure 6: Screenshot of a for-profit Telegram room for deepfake sexual exploitation with over 220,000 participants (G. Park, 2024a)

Scale and Scope

Illegal deepfake sexual crimes became rampant, with perpetrators targeting women and minors by region, university, and even personal acquaintances. Uncovered networks revealed individual chat rooms for around 70 universities nationwide divided into graduating classes, and specific rooms for distributing materials of minors per region.

Telegram groups facilitating these crimes attracted a massive number of participants. By August 2024, a single Telegram channel for the production of illegal synthetic materials was thriving with over 220,000 participants [Fig. 6], easily accessible via searches on platforms like X (formerly Twitter) (G. Park, 2024a, para. 3). Some reports cite the room as having upwards of 400,000 users (H. Lee, 2024b, para.1-4). This channel operated on a "for-profit structure" where illegal composites were made for free for the first two photos, then became a paid service (around 0.49 USD per photo in cryptocurrency), with discounts for bulk purchases or by inviting friends to expand the user base (G. Park, 2024a, para.4-5; H. Lee, 2024a, para. 15). While it was later clarified that this figure included foreign users and bots and was not solely Korean participants, it still indicates the immense scale and accessibility of these illicit activities (G. Park, 2024a, para.2; Y. Lee, 2024, para. 6-7; Ko, 2024, para. 12; C. Lim, 2024, para 1). Other groups specifically targeting middle and high school students similarly had thousands of members, with one channel involving around 2,340 participants (Go, 2024, para.6).

While exact figures are difficult to ascertain, statistics suggest that the total number of deepfake damage support cases reached 2,154 by August 2024, with some experts believing the true number of victims is likely to exceed 10,000 (Digital Watch, 2025, para. 2; J.-H. Jung, 2025, para 1-3). The crime's reach extended across a vast geographical and institutional spread, impacting over 400 schools nationwide, from elementary to prestigious universities, with lists of affected schools circulating on social media, amplifying anxiety among students, parents, and teachers (Ko, 2024, para. 12; H. Lee, 2024a, para. 3-5; H. Lee 2024b, para.1-4).

The crimes extended beyond university students to include middle, elementary, and high school students, teachers, nurses, and even female soldiers, with female teacher rooms, and female soldier rooms emerging (H. Lee, 2024a, para. 11-13; JTBC, 2024, para, 1). A ‘Deepfake Victimised School Map’ [Fig.2] exceeded 400 schools, highlighting the widespread nature of the problem (H. Lee, 2024b, para.1-4). A concerning aspect of the epidemic was the demographic of the perpetrators; roughly 73% of suspects involved in deepfake imagery and video crimes in 2023-2024 were teenagers, with around 20% in their twenties, resulting in an estimated 95.8% of suspects being minors or young adults (Oh, 2024a, para. 8; H. Lee., 2024a, para.10). In many reported cases, perpetrators were male students, and the victims were female students of the same age; a significant portion of identified online child sexual exploitation material (CSAM) in South Korea, specifically 28.1%, was found to involve deepfakes (J. Lee, 2024, para.4).

The deepfake epidemic not only revealed a quantifiable crisis but also starkly exposed a significant qualitative layer of exploitation and gendered dynamics. Victims of NCII and AIFA have expressed fear and helplessness, noting that even private social media accounts offered no true security (Go, 2024, para. 7; Y. Choi, 2024f, para. 1&3). The knowledge that acquaintances could be perpetrators caused distress; the widespread fear led many students to delete their photos from social media (Go, 2024, para. 7; S. Choi et al., 2024, para. 1-3, 5). The crisis also highlighted secondary harms, where victims faced victim-blaming and concerns about school reputation, which could silence victims from reporting (Jeong, 2024, para. 5-6; Y.Choi, 2024a, para. 3). The very act of one's photo being manipulated for sexual desire causes great mental distress and affects trust relationships (Y. Choi & Park, 2024, para.6).

Conversely, while female victims and women across South Korea responded with widespread terror and anxiety, male responses, particularly in online forums and comment sections, displayed a notable lack of remorse or empathy, directly contrasting the distress experienced by women at the time (. These included comments suggesting victims deserved the abuse, claims that victims were "not attractive enough to be deepfaked" [Fig. 7], offers of advice on how perpetrators could avoid getting caught [Fig. 8], and even direct mockery of news coverage and threats against reporters [Fig. 9]. These reactions underscored a pronounced gendered divide in the perception of this new technology and its harms, further exacerbating the distress of victims.

Figure 9: Translated screenshots from a South Korean blog post of a UOS student (from Reddit, u/Inner_Response_1714, 2024

Figure 8: Screenshot of a translated blog post detailing how to wipe evidence of deepfake crimes (from X, @KORmennow, 2024)

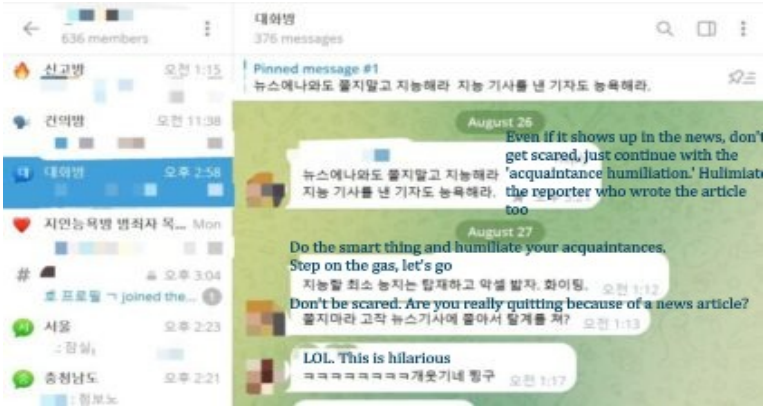

Figure 7: Screenshot showing messages from inside a Telegram deepfake groupchat (Petsenidou, 2024)

Challenges in Enforcement and Legal Loopholes

Enforcement agencies faced significant hurdles in the wake of the deepfake epidemic (H. Lee, 2024a, para 27). Notably, as with previous Nth room cases, Telegram's overseas servers complicated investigations, making it difficult for domestic agencies to request content deletion or identify distributors (Go, 2024, para. 8; J. Lim, 2024, para. 7-9). Legal loopholes, such as the requirement to prove "intent to distribute" for punishment, meant that even if illegal composite videos were produced, perpetrators could avoid charges if distribution wasn't proven (Y. Choi & Park, 2024, para. 1). Moreover, simple possession, storage, or viewing of illegal composite materials was not punishable unless the victim was a minor (J. J, Lee., 2024, para. 16-18). This resulted in lenient penalties, with most first-time offenders receiving suspended sentences or probation, rather than imprisonment, even for serious offenses (H. Lee, 2024a, para. 20-23). Overall, despite the scale, South Korea's initial institutional reactions remained performative, reflecting societal indifference to women's digital autonomy (Go, 2024, para. 9; S. Park, 2024, para 4-5). Even with revisions to the Sexual Violence Punishment Act mandating imprisonment for up to five years or fines of up to 50 million won for deepfake sex crimes, actual punishment remained lenient (H. Lee, 2024a, para. 20-23). The perceived difference in severity compared to other digital sex crimes, where deepfakes are seen as "fabricated" rather than direct exploitation, contributed to this lenient treatment (J. Lee et al., 2024, para. 5).

Responses and Calls for Action

The growing crisis prompted responses from numerous governmental bodies, educational institutions, and civil society:

Government and Police: The Ministry of Education belatedly began assessing the damage, requesting reports from provincial offices of education and urging preventive education and reporting (S. Choi et al., 2024, para. 10&15). The Seoul Metropolitan Police Agency issued ‘emergency school bells’ and distributed warnings about punishment and reporting methods. Law enforcement pledged resources, but police officials also noted the need for advanced investigative techniques due to the use of secure messengers like Telegram (S. Choi et al., 2024, para. 12; Go, 2024, para. 8-9)

Legal Reform Advocates: Experts and former prosecutors began calling for urgent measures, including deleting Telegram apps from app stores for non-cooperation, establishing ’emergency measures’ to prevent damage, and creating orders to preserve video evidence (Go, 2024, para.8; Y. Choi, 2024b, para. 4-6). There were strong calls for revising laws to punish simple possession and viewing, not just distribution, and for courts to recognise the seriousness of the crime (Go, 2024, para. 9).

Political and Social Movements: Former Emergency Response Committee, who revealed the original Nth Room incident, argued that the spread of deepfake pornography should be declared a

‘national disaster’ and criticised the government's perceived inaction (Y. Choi, 2024b, para. 1-2). Artificial intelligence experts globally signed an open letter in February 2024 - Disrupting the Deepfake Supply Chain - calling for stricter regulations, including full criminalisation of deepfake pornography (H. Lee, 2024a, 2024a, para. 28-29).

Educational Response: Educational institutions and teachers' unions called for urgent psychological counselling for affected students and teachers, and for comprehensive victim support systems (W. Lee, 2024, para. 2-4). They also emphasised prevention education to ensure students understand the serious consequences of such criminal behaviour (G. Park, 2024b, para. 6; C. Im, 2024, para. 1, 4-6). However, some educators noted that for teenagers, deepfakes might be seen as "an old, common prank," highlighting a need for substantial sex education beyond formal, brief sessions (Y. Choi, 2024d, para. 7).

Victims and activists repeatedly called for stronger measures, yet the continued spread of deepfake sexual exploitation material underscores the urgent need for comprehensive legal, educational, and technological responses to protect individuals from this evolving form of digital violence. The 2024 South Korean deepfake epidemic thus underscores the urgent need for a more robust legal and societal framework to address AI-generated sexual exploitation, demanding a fundamental shift in perception and response to protect victims and build a safer digital future.[9]

[9] All references can be located in the main resource list above

10. Appendix B; Technical Overview of Deepfake Generation

This appendix provides a detailed technical overview of deepfake generation technologies, emphasising the underlying machine learning architectures and computational processes that enable the creation of highly realistic NCII. This information serves to complement the theoretical discussion in Chapter 2 by offering a deeper insight into the *how* of deepfake creation. Deepfakes, as discussed throughout Chapter 2, are a sophisticated form of synthetic media that leverages advanced artificial intelligence and machine learning techniques to manipulate or generate visual and audio content. These technologies are capable of superimposing or altering human likenesses with unprecedented realism, challenging established paradigms of identity and perception.

To achieve this, the technology relies on two primary machine learning architectures: Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) (Mahmud & Sharmin, 2020, p.2):

Generative Adversarial Networks (GANs)

GANs comprise two neural networks - Generator network (G) and a Discriminator network (D) - engaged in a continuous adversarial game: the Generator attempts to create indistinguishable synthetic content (e.g., a fake image or video frame), while the Discriminator learns to differentiate between real and fake content (Brooks et al., 2021, p. 12). This iterative competition drives both networks to improve, with the Generator becoming increasingly adept at producing highly realistic fakes that can deceive even advanced detection algorithms; the adversarial nature of GANs is crucial for achieving the high fidelity and realism seen in modern deepfakes.

Variational Auto-Encoders (VAEs)

Alongside this, VAEs use an encoder to compresses input data (e.g., a person's face) into a lower-dimensional latent space via an encoder network, and a decoder that reconstructs the image back into the original input from this compressed representation via a decoder network (Somers, 2020, para 13-18). In the context of deepfakes, a shared encoder is often trained on multiple faces, common facial features (Somers, 2020, para 13-18). By feeding the latent representation of one person's face through the decoder trained on another's face, the system can effectively swap identities while preserving expressions and movements, allowing for the creation of new, non-consensual imagery by combining elements from different individuals (Harris, 2019, p.100). VAEs are particularly effective for face swapping due to their ability to disentangle and recombine facial attributes within the latent space.

The Deepfake Generation Pipeline

These neural network systems require algorithmic training with multiple video samples capturing comprehensive facial characteristics [fig.10]. The deepfake pipeline follows several stages. It begins with *Data Acquisition and Preprocessing*, where extensive image or video datasets of both source (target's face) and target individuals (person whose face will be replaced) are gathered

followed by face detection, landmark alignment, and normalisation (Harris, 2019, p.100-101; Mahmud & Sharmin, 2020, p.15-16). The quality and diversity of this foundational dataset are paramount for the realism of the final deepfake. This data fuels the *Model Training* phase, during which GANs or VAEs are trained over hundreds of thousands of iterations, teaching the AI to mimic an individual's unique facial characteristics, expressions, and speech patterns (Brooks et al., 2021, p.12&27; Mahmud & Sharmin, 2020, p.15-16). The iterative nature of machine learning means that the neural networks continuously refine their capabilities, learning from each generated output to improve realism. Finally, the actual *Face Swapping/Generation* applies these trained models to new source and target content to produce altered media (Brooks et al., 2021, p.27). This involves the AI generating new facial content for the target, seamlessly integrated with their body and environment. This core process culminates in *Post-Processing* - a key step for achieving realism that involves blending the synthesised face, colour correction, lighting adjustments, and artefact reduction, all contributing to the creation of seamless digital fabrications of an individual's image (Mahmud & Sharmin, 2020, p.17). Recent research identifies effective deepfake generation now only necessitates a 30 second video sample capturing a variety of facial angles, expressions, and diverse lighting conditions to generate a sufficiently comprehensive dataset (Wrona, 2025, para. 1&7). Crucially, this foundational data is frequently acquired without the explicit consent or knowledge of the individual depicted, laying the groundwork for subsequent non-consensual manipulation, also understood as NCII (Non-Consensual Intimate Images) (Brooks et al., 2021, p.17; Kobis et al., 2024, p.2).

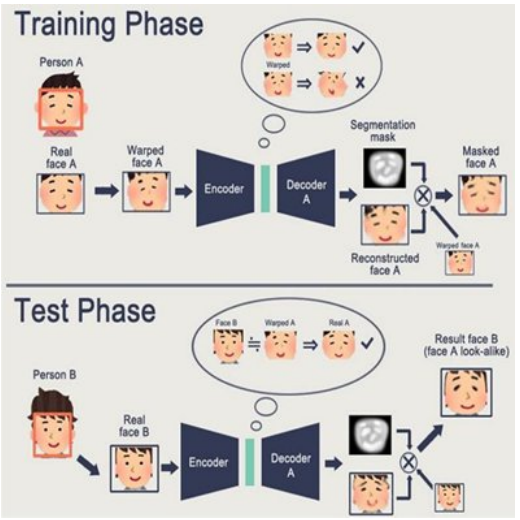

Figure 10: Illustration from Mahmud & Sharmin (2020, p.17) detailing the training and testing phase of deepfake models

Primary Deepfake Methodologies

The methodology of the deepfake process relies on two main approaches [fig.11]: *Face swapping*, which involves digitally transplanting facial characteristics by analysing and replacing specific facial regions (eyes, nose, and mouth) while maintaining the facial contour and skin tone and; *Facial manipulation*, which takes a more nuanced approach, by synthetically reconstructing

facial expressions to alter an individual's perceived emotional state or verbal communication (Howard, 2020, para, 3-4; Mahmud & Sharmin, 2020, p.15-17; Brooks et al., 2021, p.10). Technical finesse in both methodologies involves the neural network's ability to analyse subtle facial characteristics, meticulously mapping facial landmarks, understanding skin tone variations, and reproducing complex facial movements to generate seemingly real, recreating diverse facial angles, expressions, and lighting conditions to do so (Brooks et al., 2021, p. 13-14). In tandem with this, the iterative nature of machine learning means that deepfake technologies become increasingly performant with each use, whereby the neural networks continuously refine their capabilities, learning from the generated content (Brooks et al., 2021, p.5). Consequently, this process of constant improvement progressively erodes the reliability of human perception, with the U.S. Government Accountability Office (2024) forecasting that distinguishing computer-generated content from authentic media will become a formidable challenge as the technology matures (Obadia, 2024, para. 6).

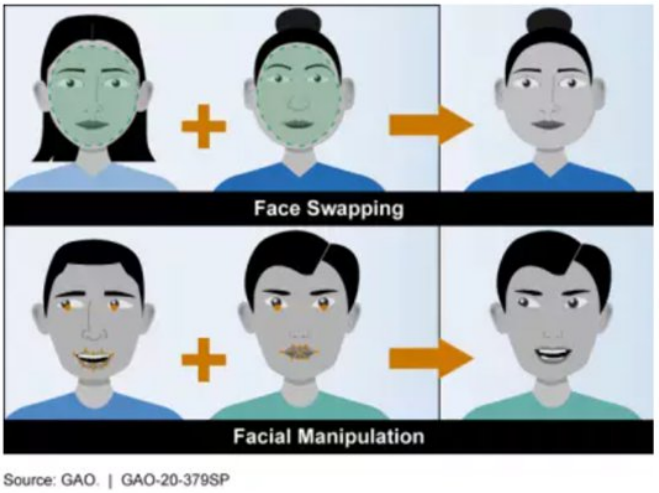

Figure 11: Illustration from Morgan (2020) showing the difference between facial swapping and facial manipulation

Technological Evolution and Detection Challenges

This technological evolution of deepfakes has significantly eroded the reliability of human perception: a 2025 study revealed that only 0.1% [of 2000 participants] accurately distinguish real from deepfaked content (Charlwood, 2025, para 2). Unlike traditional image manipulation, AI systems can now compensate for previous limitations; where earlier iterations initially struggled with complex features like eyes, ears, and hands, current versions of AI deepfake technology can generate uncanny representations by understanding subtle contextual cues and facial nuances, 'learning' from mistakes (Mahmud & Sharmin, 2020, p.16-18).

Compounding controversies, deepfake accessibility and democratisation has accelerated through user-friendly open-source software platforms like DeepFaceLab, Faceswap, ReFace, and Zao that offer interfaces requiring minimal technical expertise (Brooks et al., 2021, p. 9; Kobis et al., 2024, p.2). Tools like DownAlbum and Instagram Scraper also enable users to download entire

image collections from social media platforms, providing the foundational dataset for deepfake creation (Harris, 2019, p.101). [10]

[10] All references can be located in the main resource list above

11. Appendix C – Comprehensive Discussion and Analysis of Findings

This appendix contains the full and detailed discussion and analysis of the research findings, which comprehensively addresses the primary research question: How are the impacts and perceptions of deepfake technologies related to the underrepresentation of women in AI development and regulation? It provides the complete empirical evidence, theoretical application, and interpretative depth that underpin the conclusions presented in Chapter 6 of the main body.

C. I Underrepresentation's Enduring Imprint on Deepfake Realities

The issue of AIG-NCII, and the gendered digital violence it inflicts, demands a deeper examination. While the preceding chapters outlined the empirical realities, the following discussion interprets these findings through the previously established theoretical lens to directly address the primary research question: *How are the impacts and perceptions of deepfake technologies related to the underrepresentation of women in AI development and regulation?* Findings suggest that the unique and disproportionate harms women endure are not coincidences, but systemic consequences of a technological landscape fundamentally shaped by an underrepresentation of female voices. Indeed, these systems, shaped without centring women's experiences and understanding the specific vulnerabilities digital technologies create - evidenced in diversity statistics regarding AI development and regulation - may fail to adapt to the nuanced challenges of this new form of gendered violence (Zowghi & Bano, 2024, p.3-4; Custer & Fosch-Villaronge, 202, p.72, 80-81, Aavik et al., 2024, p.3). To that end, this chapter argues that truly addressing these violations necessitates centring women's experiences in AI development and regulation. This imperative, central to the emergent Digital Violation Discrepancy theory, highlights how differing understandings of deepfake harms, influenced by gender and generational perspectives, perpetuate challenges within patriarchal technological systems.

C. II The Perennial Chasm: Why Consent Violations Go Unaddressed

The dynamics of sexual violation in online spaces often mirror, and even amplify, the ambiguities inherent in real-world interpretations of harm, particularly concerning consent (Killean et al., 2022, p.855). Historically, societal and legal discourses around sexual assault have frequently grappled with defining consent, often prioritising overt physical acts or external indicators over the subjective experience of violation, leading to uncertainties surrounding culpability and victim experience (Randall, 2010, p.422-426; Guajardo & Tadros, 2024, p.24-25). Deepfakes thus emerge within this pre-existing societal pattern of navigating consent's complexities that deepfake technologies have emerged, translating these ambiguities into the digital realm with even greater uncertainty and consequence.

The disproportionate impact of deepfake harms, particularly on women, appears to be influenced by the underrepresentation of female voices in AI development and regulation. As female focus group participants revealed, the generation of non-consensual deepfakes was consistently highlighted as the foundational act of harm, an irrefutable violation of personal autonomy regardless of dissemination. This understanding, emerging from a marginalised standpoint, critically challenges conventional interpretations of digital harm, which often prioritise visible, externally measurable consequences over the violation of one's digital self (Lageson, 2016, p.23-25). The viewpoint resonates strongly with Butler's (1999, p.29-30) gender performativity concept, where deepfakes can be seen to impose a fabricated digital performance onto an individual, thus assaulting their agency and bodily integrity at the very point of creation. Women acutely perceive and vocalise this initial violation; their lived experiences of historic marginalisation and exclusion from the male-dominated creation of these technologies may render them uniquely attuned to the subtle yet impactful forms of digital objectification. Centring their experiences offers a vital epistemological lens, revealing deepfake harm begins not with public exposure, but with the invasive digital imposition of a fabricated self.

In contrast to this, male discourse tends to prioritise distribution as the primary source of harm, focusing on visible, external, or legal consequences rather than the intimate violation of consent. This externalised view of harm appears to be a consequence of AI being predominantly developed and governed by a homogeneous male creator base, in which their perspective, less attuned to the subtle violations of digital bodily autonomy, focuses on tangible, publicly observable outcomes, thus potentially overlooking the initial, intimate violation for women (Zowghi & Bano, 2024, p.3-4; Custer & Fosch-Villaronge, 202, p.72, 80-81; McKinnon et al., 2024, p.88-91). This male-centric lens, seemingly embedded within AI development, could contribute to the exclusive harms women face by ensuring that the precursory act of digital violation remains largely unaddressed (Custer & Fosch-Villaronge, 202, p.72, 80-81; Aavik et al., 2024, p.3; Vazquez, 2021, p.19; T.Williams, 2024, para. 1, 8-12). This societal bias reflected in Western [US/UK] media narratives, which largely emphasise distribution-based harms, frames deepfake crimes as sensational traumatic incidents rather than explicitly illegal acts at the point of creation. This widespread media framing, interpretable as biases spilling over from underrepresentation in the societal understanding of these technologies, reinforces this male standpoint, downplaying the inherent consent violation in deepfake creation for women. Such a systemic oversight may ensure that the harms for women stemming from this initial violation remains largely unacknowledged in public discourse and subsequent policy.

This Digital Violation Discrepancy in perceiving harm suggests significant implications for regulatory responses, particularly a miscalibration in their design. Current efforts, exemplified by the US and UK's approaches, largely maintain a traditional focus on punishment for dissemination rather than prevention at the source (Mullin, 2025, para. 1-2; Narayanan, 2024, para. 13-14). This punitive

emphasis aligns deeply with a long-standing societal and political philosophy which posits that a strong legal framework and the threat of severe consequences are the foundation for enforcing social expectations and maintaining order, serving as the primary bulwarks against a descent into chaos (Hobbes, 2009, p. 86, 116, 142-147, 162-164). For this deterrence model to succeed, the exposure of wrongdoing (evidence) and the certain imposition of consequences (punitive laws) are considered essential. In theory, in the online space, with its somewhat-ubiquitous Foucauldian "surveillance state", characterised by the modern development of data assemblages and data doubles, may, theoretically, enhance the efficacy of punishment by making digital actions traceable and accountable (Galič et al., 2016; pp. 10; Foucault, 1995; pp.207-208 & 215; Haggerty & Ericson, 2003, p.606). The omnipresence of digital footprints – from IP addresses to usage logs – ostensibly suggests a landscape ripe for traditional judicial oversight, where threat of legal repercussions should prevent misconduct (Lageson, 2016, p.23-25).

Diverging from this expectation, deepfakes introduce a new challenge to the traditional punitive paradigm, especially regarding harms inflicted upon women. The anonymity afforded to perpetrators - particularly through privacy-forward networks like Telegram and Signal where the main distribution appears to take place - fundamentally undermines the deterrent effect of punishment (Reuters, 2024, para.4; H.Park, 2024, para 2&6; J.-J. Lee, 2025, para. 5-8). These platforms are explicitly designed to encrypt communications and minimise user data retention, creating something akin to digital havens where traditional surveillance and evidence collection techniques struggle to penetrate. This architectural design, while purportedly serving user privacy, simultaneously enables a culture of impunity for digital abuse. As illustrated by the emergence of *digital undertakers [디지털 장의사]*, offering illicit services to erase digital footprints and evidence of deepfake creation and dissemination, AIFA perpetrators seem acutely aware of, and actively exploiting, these loopholes in traditional judicial systems, understanding that, without tangible evidence, traditional courts struggle to build and prosecute a case, effectively flaunting technological affordances that may grant them perpetual impunity (Yang, 2024, para. 6-7, 11-12; I. Im, 2024, para.1-10). This suggests that punishment, while a necessary component of justice for the visible harms, cannot be the sole or primary preventive measure - its efficacy as a deterrent appears severely hampered when the act can be easily circumvented and remain untraceable.

This deep-rooted disconnect between existing legal frameworks (designed for a pre-digital or less complex digital era) and the realities of deepfake technology – particularly its sui generis anonymity – contributes to the persistence of harms for women and an inability to tackle the increasing issue of deepfakes. This failure is further entrenched by the media's disparate framing across regions, where the Western (US/UK) narratives, as highlighted in the CDA, consistently emphasise the horrifying *outcome* over the illegal nature of the *initial act* (point of creation), or the systemic failures enabling impunity. In contrast, South Korea's experience, marked by a high volume of consistent public discourse that explicitly frames deepfake crimes as illegal, suggests a more

effective pathway to legal redefinition, directly linking societal dialogue to comprehensive regulatory responses that can begin to address the root cause of these violations and better protect digital autonomy.

C. III The Unaddressed Ethical Void: Underrepresentation and Platform Complicity

Consistent throughout, the findings point to yet another paradox within this deepfake regulation issue: the online entities that govern digital interaction, while frequently implicated in the facilitation of these crimes, have been, and continue to be, rarely held directly accountable for their role. These actors -broadly conceptualised as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) aligning with existing legislation under the EU's Digital Services Act – are core architects of the digital environment and central to the unchecked flow of digital services that enable AIG-NCII creation, dissemination, and visibility (European Commission, 2022, Article.33; Ding & Suresh, 2025, para.5-9). Yet, their operational design often hints at a lack of ethical foresight concerning these new digital harms; this absence of proactive considerations and corresponding legislative failure to anticipate or adequately address these digital harms, despite media coverage emphasising impacts and their implicit role, directly contributes to the harms women experience through deepfake exploitation (Ding & Suresh, 2025, para.5-9; Plant, 1997, p.38-41). Furthermore, the very nature of these VLOPs and VLOSEs as detached, faceless entities - vast, seemingly impersonal corporations operating at a global scale - fosters a societal perception that they are somehow beyond effective human control or conventional regulation, exacerbating their impunity (Crofts, 2024, p.375-377).

This ingrained lack of accountability fuels the architecture of exploitation and informatics of domination; platforms, often engineered without a deeply diverse ethical compass, lack the safeguards to mitigate harms disproportionately affecting groups outside their creator demographic (Custer & Fosch-Villaronge, 2022, p.92, 239, 28; Ramos, 2022, para. 9; S. Pal et al., 2024, para. 2; Haraway, 2016, p.27-31). This systemic issue is compounded by media narratives, which, as the CDA identified, showing a near-unanimous absence of explicit calls for platform responsibility in media narratives, despite nearly every article implicitly acknowledging the role of online entities as primary mediums for deepfake creation and distribution. This unaddressed culpability is particularly concerning when considering the continuum of harm facilitated by these entities.

VLOSEs, for instance, are deeply implicated in enabling the creation of AIG-NCII; the continued allowance of user-friendly deepfake tools and marketplaces (e.g. Clothoff, Nudify, and DrawNudes) is a direct manifestation of this unchecked development (Gulsen & van der Plas, 2025, para.1; Koltai et al., 2024, para 4, 5, 12; Higgins et al., 2025, para. 24). Studies indicate at least 41 such platforms operating openly on the Clearnet, easily discoverable via common search engines, affirming this regulatory failure (Singh, 2024, para.3-6; KeyNorth Group, 2023 para.2). This widespread accessibility, implicitly highlighted in media describing "easily accessible apps", begs

the question of why these services are permitted to exist – a sentiment echoed by the female focus group participants, who consistently argued that such services "shouldn't even be possible in the first place," necessitating regulation that prevents their very operation.

This leads to a Pandora's Box-type dilemma when understanding how to regulate the creation of deepfakes, emphasised by male focus group participants whose conceptualisations of tech regulation were often accompanied by excuses for its futility, acting as a convenient rhetorical shield against accountability. In that essence, it is acknowledged that, given that deepfake technology is irrevocably 'out there', it may indeed be impossible to entirely prevent a determined individual from writing their own code, training a private model, and operating on a local, isolated network. In such hypothetical, self-contained instances, without distribution, the afforded anonymity of this crime could, indeed, suggest perpetual impunity. However, this inherent issue in achieving absolute private containment does not, and philosophically cannot, justify the egregious ease with which these tools and their harmful outputs remain accessible on the broader public web. Ergo, the moral imperative demands a clear distinction: the impossibility of absolute, private suppression does not excuse the societal negligence of allowing rampant public access. The implicit suggestion in media representations, implying ease of creation and ability to inflict widespread harm quickly, especially if minors can do it on such a scale, underscores the urgent need to curb this accessibility.

Hand-in-hand with this, VLOPs, too, bear significant responsibility for distribution and encountering AIG-NCII - a role frequently highlighted in the media representations. While distribution often receives more legislative focus, its regulation remains questionable. For instance, the US's Take It Down Act mandates VLOPs remove deepfake content within 48 hours, however, this timeframe appears performative, neglecting the virality of digital content; intimate imagery can be downloaded, re-shared, and re-uploaded countless times within mere hours, rendering post-hoc removal efforts largely ineffective and offering little recourse for victims (Ortutay, 2025, para 3; Mullin, 2025, para. 1-2). This contrasts with the removal policies implemented in South Korea, where flagged content is removed within 24 hours, pending subsequent review, and platforms are encouraged to block suspicious content pre-emptively and conduct reviews subsequently; this ‘act first, question later’ approach appears to be a more effective response in limiting the immediate and subsequent harms of distribution (Seok et al., 2025, p.3-4; S.Park et al., 2024, para.3-4).

Furthermore, the very nature of privacy-forward messaging applications, such as Telegram and Signal, which, while offering important affordances for secure communication, are also the primary sites of the large-scale distribution networks that are being uncovered, demanding direct and proactive cooperation from regulators and platforms alike (H.Park, 2024, para 2&6; J.-J. Lee, 2025, para. 5-8; S. Park et al., 2024, para.5). This deliberate misuse of privacy-centric design necessitates that these platforms be regulated to cooperate with law enforcement when deepfake content is identified, a critical issue highlighted by Telegram's lack of cooperation in the South Korean deepfake epidemic (Go, 2024, para.8; Uren, 2025, para. 15).

This moral imperative reflects the gendered split observed in the focus groups; while the female focus group participants consistently sought regulatory solutions for platforms and accessibility despite perceived challenges, male participants often expressed a resignation to the inevitability of circumvention, echoing a 'why bother' sentiment. This male-centric viewpoint, perhaps embracing the notion of VLOPs and VLOSEs as insurmountable entities beyond human control, not only undermines efforts to limit harmful digital flows but implicitly tolerates the very mechanisms of digital exploitation that exist on the everyday web. Such resignation, reflecting a collective societal negligence, allows the ethical void to persist, directly enabling the unchecked proliferation of deepfake harms.

C. IV Normalising Digital Violence: Underrepresentation and the Perversions of Minor's Realities

Consistent throughout the discourse surrounding deepfakes, there is the recognised, yet often underexplored, paradox concerning the involvement of minors: their prevailing roles as both vulnerable victims and a significant demographic of perpetrators. While this dual reality is statistically acknowledged, the depth and nuance of its implications for policy and societal response remain largely unaddressed, highlighting a critical gap in our understanding of digital harm, particularly concerning young individuals.

A critical, yet under-discussed, element of this problem manifests in attempts to regulate victimhood, specifically concerning minors and sexual violation. This is not a new issue; rather, it is a continuation of a pre-existing societal pattern where the sexualisation of minors has, to varying degrees, been normalised or subtly excused within cultural discourse (Papadopoulos, 2010, p.33-52). Consider the widespread media fascination surrounding Emma Watson's 18th birthday, or the public debate ignited by Bhad Bhabie's reported million-dollar earnings on explicit OnlyFans only 6 hours after turning 18 (Sayer, 2016, para.4; Young, 2021, para.1). These examples underscore a collective discomfort with truly defining and enforcing boundaries around consent and exploitation for young individuals.

This underlying societal ambiguity around minor sexualisation finds itself built into arguments used to defend certain cultural products that further normalise minor sexualisation. Take, for instance, the justifications used to depict lolicon - a portmanteau of Lolita Complex referring to a genre of Japanese popular culture (primarily manga, anime, and artwork) that depicts young or young-looking female characters in a sexually suggestive or erotic manner - argue it is merely artistic fantasy, and that the exclusion of a *real* minor detaches it from *real* harm (Takeuchi, 2015, p.198, 210; McLelland, 2005, p.4). With the emergence of AIG-CSAM, similar arguments are now being repurposed: the contention posits that because AIG-CSAM involves a synthetic or fabricated child's body, it supposedly allows paedophilic urges to be vented without causing *real* harm [Fig. 12] (NCMEC, 2025, para.3-4). This argument, while superficially offering a containment strategy for

illicit desires and protection for other vulnerabilities minors face (paedophiles and online grooming), ultimately sidesteps the harm true nature. Thus far, deepfake analyses suggest that harm extends far beyond the physical and tangible, encompassing the initial violation of consent and the digital assault on identity; ergo, claiming "real harm" is absent when digital autonomy is violated constitutes a significant conceptual misstep. Such rationalisation reflects a larger societal construct that, for inexplicable reasons, continues to be deployed to defend the sexualisation of minors, and while no formal link has been found, experts often argue that exposure to such material actively encourages and normalises paedophilic behaviours, leading to inverse effects and increasing the likelihood of real-world harm (Takeuchi, 2015, p.230). This line of reasoning forms the foundation preventing effective regulation of AIG-CSAM in countries like Japan, leaving countless minors uniquely vulnerable to pervasive digital exploitation (UNICRI, 2024, p.27).

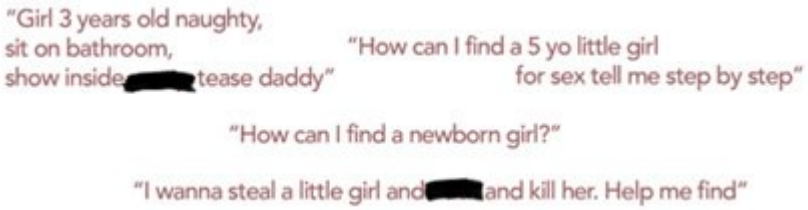

Figure 12: Real prompts used to create AIG-CSAM [Missing Kids - https://www.missingkids.org/blog/2024/generative-ai-csam-is-csam]

Mirroring the struggles of regulating victimhood, the challenge of addressing minor perpetration similarly exposes societal blind spots and biases. As the CDA revealed, Western [US/UK] media representations are more hesitant to explicitly frame minor deepfake involvement as a crime, instead prioritising the traumatic victim experience and implicitly adopting an attitude that suggests "they didn't know better." Such framing, which often portrays these acts as unintentional misdemeanours rather than serious crimes, not only neglects the underlying systemic factors motivating perpetration, but also reflects a digital immigrant perspective on the realities of online interaction. This perspective frequently struggles to reconcile traditional notions of childhood innocence with the realities of digital malice that are present today, overlooking that experts contend that minors involved in these acts are aware of their illicit nature, engaging in them as more than just a joke (Thorn, 2025a, p.24-27). This broader societal lack of understanding, particularly concerning AI's fundamental workings, likely contributes to the misinterpretation of minors' roles in perpetration in the West, potentially leading to slower or limited regulation (Heikkilä, 2024, para.1-4).

In direct contrast, South Korea, currently grappling with the aftermath of its nationwide deepfake epidemic, has been compelled to confront this reality directly. Its legal and societal response has been swift and severe, transitioning from virtually no specific punishment for minors involved in deepfake crimes to an inescapable legal and judicial response for any minor caught creating *or* distributing AIG-NCII (Soek et al., 2025, p.2-9; C. Kim, 2024, para. 3-4). This punitive

shift is mirrored in its media representation, which consistently reflects the high statistics of minor involvement (73% minor perpetrators; 28.1% minor victims), keeping the issue front and central in public discourse (J. Lee, 2024, para.4; Oh, 2024**a**, para. 8). However, even the most stringent punitive measures face limitations, particularly given the sui-generis affordances of anonymity. The emergence of digital undertaker services, which is seen to be exploited by minors in South Korea to erase digital footprints and evade traditional evidence collection, underscores that perpetrators, even minors, are acutely aware of the need for evidence in judicial systems and follow through of these punitive measures (Yang, 2024, para. 6-7). This raises a critical question: is punishment alone truly enough when the very anonymity and circumvention capabilities of deepfakes that we are seeing can hamper its enforcement? This challenge suggests that while punitive measures serve as an important deterrent, highlighting that perpetrators are being held accountable to curb such behaviour, they are insufficient in isolation; the female focus group participants consistently advocated for a more holistic approach for minors specifically, emphasising rehabilitation and re-education, with focus on social aspects like understanding consent online, and starting as early as "sharing is caring", rather than relying solely on punitive measures. This perspective underscores the need for a tandem approach: punitive deterrence, even when hampered by anonymity, must be coupled with fundamental education to complement and rehabilitate behaviours and understandings. This integrated strategy, prioritising both accountability and ethical development, is critical for comprehensively addressing deepfake harms and protecting the most vulnerable in the digital age.[11]

[11] All references can be located in the main resource list above

12. Appendix D - Policy White Paper: Recommendations for a Victim-Centred Approach to Deepfake Regulation

Based on Insights from “Behind the Screens: Gendered and Generational Divides in Understanding Deepfake Violence”

Anja Ellwood
MA Digitalisation, Surveillance & Societies, Erasmus University Rotterdam
June 24, 2025

Summary

Deepfake Non-Consensual Intimate Imagery (NCII) represents an escalating threat, disproportionately impacting women and minors. Current regulatory and technological responses are proving inadequate, enabled by a Digital Violation Discrepancy that appears to prioritise technological advancement over victim safety and consent. This white paper, drawing on the findings from the thesis "Behind the Screens: Gendered and Generational Divides in Understanding Deepfake Violence", proposes a victim-centred regulatory framework for deepfake content. The following outlines seven recommendations for policymakers, technology developers, educators, and legal systems to proactively address the unique harms inflicted and encourage a safer digital environment. Implementing these recommendations will foster more equitable technology development, establish robust preventative safeguards, enforce platform accountability, enhance digital literacy, and ensure empathetic and effective responses, ultimately cultivating a digital environment that prioritises consent, identity, and bodily autonomy, empowering individuals and protecting vulnerable populations.

Why Current Approaches Fall Short

Current responses to this escalating threat often fall short, characterised by reactive measures and an inherent Digital Violation Discrepancy. This discrepancy highlights a gap where the speed of AI development, frequently driven by homogeneous and uncritical perspectives, creates new avenues for violation that existing ethical considerations, legal frameworks, and protective mechanisms fail to comprehensively address. The result is a digital landscape where the informatics of domination persists, allowing technologies to exploit rather than empower, especially for marginalised groups such as women and minors. This white paper thus outlines a comprehensive victim-centred approach to rectify this imbalance and cultivate a safer, more equitable digital future.

The pervasive nature of deepfake harms is rooted in systemic issues that transcend individual malicious acts; the ethical void identified in current regulatory and developmental practices stems from several factors:

Homogeneous Creator Bases: A significant lack of diversity within AI development teams and ethical review boards perpetuates underlying biases. AI technologies are often built without a holistic understanding of potential harms to diverse user groups, leading to oversight of vulnerabilities experienced by women and minors. This patriarchal bias, as identified in the theoretical framework, results in systems that inadvertently or directly facilitate harm.

Reactive Remediation: Existing policies and platform responses largely focus on post-hoc removal of content rather than proactive prevention. This reactive stance places an immense burden on victims, who must contend with the initial violation and the arduous process of reporting and removal, often after widespread dissemination has occurred. The emphasis on ex-post facto measures fails to address the foundational harm: the very act of creation.

Insufficient Platform Accountability: Despite their power and capacity, Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) often exhibit a level of complicity, either through inadequate moderation, slow response times, or a lack of transparent reporting. The burden of policing harmful content should not solely rest on individual users or victims.

Gaps in Digital Literacy and Consent Education: It appears that a significant portion of the population, both digital natives and immigrants, lack comprehensive understanding of online consent, critical media literacy, and the psychological impacts of digital harms. This deficit contributes to the normalisation of harmful online behaviours and inadequate self-protection.

Inadequate Responses to Minor Perpetrators: Current punitive approaches often fail to address the underlying societal factors, developmental contexts, and peer influences that contribute to minors engaging in perpetration. A lack of restorative justice and educational interventions means cycles of harm may persist.

Exclusion of Victim Voices: A persistent oversight in much of current policy development and technological design is the inadequate centring of victims' lived experiences and trauma. As evidenced by insights drawn from this research, without direct input from those most affected, solutions risk being ineffective, insensitive, or misaligned with genuine needs.

Fragmented Global Responses: The transnational nature of deepfake creation and dissemination is met with disparate national legal frameworks and enforcement efforts, creating jurisdictional loopholes that bad actors exploit. This fragmented approach allows the "Hydra-like" phenomenon of online crime to thrive.

Recommendations for a Victim-Centred Regulation Effort

Addressing these foundational challenges requires a paradigm shift towards a victim-centred framework, integrating preventative measures, robust accountability, and holistic support.

Based on the insights provided by this research, particularly the Digital Violation Discrepancy theory and the need to centre women's experiences, this white paper posits the following practical applications and recommendations for policymakers, technology developers, educators, and legal systems:

1. Mandating Diversity and Inclusion in AI Development and Regulation: To address the root cause of the Digital Violation Discrepancy, concrete steps should be implemented to ensure diverse representation across all stages of AI development, ethical committees, and regulatory bodies. This could include setting clear targets for gender, racial, and socio-economic diversity within engineering teams, research labs, and policy-making panels. Organisations should be incentivised or legally mandated to report on their diversity metrics. By embedding diverse perspectives from inception, technologies will be built with a more inclusive understanding of potential harms, moving beyond the patriarchal biases identified in the theoretical framework. This proactive measure will inherently consider women's experiences, preventing the unique harms that arise from a homogeneous creator base.

2. Prioritising Generation-Based Prevention and Ethical Design: Policymakers should advocate for "ethics by design" and "privacy by design" principles to be integrated into AI development from the very outset. This means compelling AI developers to implement technical safeguards that actively prevent the creation of non-consensual deepfakes, rather than solely focusing on post-hoc removal. This could include developing robust algorithms that detect and block attempts to manipulate human likenesses without explicit, verifiable consent, particularly for sexual content. The focus group findings, where women consistently emphasised the generative act as the primary violation of consent, underscore the urgency of this approach. This proactive prevention addresses the unaddressed ethical void identified throughout the analyses, moving beyond reactive measures to tackle the foundational harm at its origin.

3. Strengthening Proactive Platform Accountability: Stricter regulations are urgently needed for VLOPs and VLOSEs to ensure they proactively prevent both the creation and widespread distribution of deepfake NCII. This could manifest as mandatory real-time content scanning, immediate content removal upon flagging (faster than the current 48-hour windows seen in some Western legislation), and transparent reporting on deepfake incidents and moderation efforts. Moreover, clear and severe penalties for non-compliance must be established, shifting the burden of responsibility onto the platforms that facilitate these harms. This addresses the ethical void created by platform complicity and leverages a more diverse understanding of systemic harm, ensuring that digital spaces are not enabling the informatics of domination.

4. Implementing Comprehensive Digital Literacy and Critical Consent Education: Educational institutions, supported by governmental initiatives, should look to integrate comprehensive digital literacy, consent education, and ethical reasoning into school curricula from an early age, ideally at the "sharing is caring" stage. This education should go beyond technical skills to encompass critical media literacy, the psychological and social impacts of digital harms, and a nuanced understanding of online consent. Crucially, these curricula ought to be developed and delivered by experts from diverse backgrounds who understand both technological realities and youth development. This approach will empower digital natives to critically navigate online environments and help bridge the digital immigrant gap among older generations, thereby mitigating the normalisation of harm and addressing the complex dual role of minors.

5. Adopting Nuanced Approaches for Minor Perpetrators: Responses to minor perpetrators should aim to prioritise rehabilitation and education over purely punitive measures. While accountability is necessary, policies must acknowledge minors' developmental contexts and address the societal factors (such as the normalisation of harm, peer influence, and the allure of online anonymity) that enable their actions. Restorative justice programs, counselling, and educational interventions focused on empathy, critical thinking, and the real-world consequences of digital actions should be emphasised. This approach, advocated for by female focus group participants, ensures a response that respects the complexities of the issue while still effectively addressing the unique harms for women and girls by disrupting cycles of perpetration.

6. Centring the Victim's Experience: All policy development, technological design, and support services should shift towards being informed by and prioritising the lived experiences and trauma of victims, particularly women and minors. Their voices, currently underrepresented in policymaking and development, must be central to fully addressing the harms they endure. This could include funding for accessible psychological support, comprehensive legal aid, safe reporting mechanisms, and survivor-led advocacy initiatives. Centring victims ensures that solutions are genuinely effective, empathetic, and directly responsive to the challenges deepfakes pose to identity, consent, and bodily autonomy.

7. Fostering Global Collaboration and Harmonised Legal Frameworks: Given the transnational nature of deepfake creation and dissemination, a move towards increased international cooperation is paramount. Governments and international bodies should aim to work towards harmonising legal frameworks, sharing best practices, and coordinating enforcement efforts to combat the "Hydra-like" phenomenon of online crime. This global collaboration is key to

overcoming national limitations and establishing a unified, effective response to this escalating digital threat.

Conclusion

The insights garnered from this research unequivocally demonstrate that the current trajectory of deepfake proliferation, unchecked by adequate regulation and a victim-centred perspective, poses a serious threat to individual [digital] autonomy, particularly for women and minors. Addressing underrepresentation in AI development and regulation is not merely a matter of fairness or social justice; it is an imperative for building truly ethical, safe, and equitable digital futures for all. Only by challenging the informatics of domination and actively reshaping the digital landscape to genuinely empower rather than exploit can we hope to fully address the unique harms inflicted upon women and minors by centring their experiences. As such, this white paper urges policymakers, technology leaders, and legal systems to adopt these recommendations without delay, nurturing a digital society where innovation serves humanity, not undermines it.

13. Appendix E – Focus Group Question Guides

Focus Group Question Guide: Female Group

Introduction & Ground Rules: (Approx. 10 minutes)

- Welcome participants and thank them for their time.
    - Hi everyone, thanks so much for taking the time to be here today. My name is { X }, and my colleague { note taker or helper } and I are conducting this discussion as part of a research project exploring how people perceive new technologies. Your honest opinions and experiences are really valuable to us."
- Briefly explain the purpose of the focus group (understanding perceptions of new technologies).
- Reiterate confidentiality and anonymity.
- Explain the format (open discussion, no right or wrong answers).
- Ask participants to briefly introduce themselves (name and study/faculty is all thats needed).
- "To start, could you briefly share your first name and maybe one thing you've heard or read about AI in the news recently?" (This can help ease participants in, but skip if time is tight).
- Understanding the Technology: "How familiar are you with how AI deepfakes are made? What's your general understanding of the technology involved?"
- Perceived Accessibility: "How easy do you think it is for the average person, including teenagers, to encounter deepfakes online? What about the ability to create them?" (Probe for awareness of user-friendly AI tools, online platforms.)

Perceptions of Harm: (Approx. 25-35 minutes)

- Opening Question: "When you hear the term 'AI-generated deepfake,' what are your initial thoughts or feelings? What comes to mind?"
    - Probe: can you tell us more about why this is your initial reaction?
- *Explain a bit about what deepfake technologies to make sure everyone has a base understanding*
    - *"To be written – Just to quickly ensure we're all on the same page, 'deepfakes' are realistic-looking digital creations where someone's image or likeness has been manipulated using AI, often to create fabricated videos or images."*
- Scenario 1 (General Adult Target - Initial Reaction): "Imagine you see a realistic, sexually explicit video of an adult is created using AI without their consent being shared online. What is your inital reaction to this scenario?"
    - General Probe: "What are the first things that come to mind when you picture this happening?" (Keep intentionally Gender Neutral to see if they already assume a victim's gender"
- Scenario 2 (Adult Female Target) - if vitcim one remains gender neutral): "Imagine a situation where a realistic, sexually explicit video of an adult woman is created using AI

without her consent and shared online. What kind of harm do you think this could cause her?" (Probe for different types of harm: emotional, psychological, social, reputational, safety, etc.)

  - If the initial responses are general/ if the initial victim is gendered:
  - Revised Probe for Emotional/Psychological Impact: "Could you tell us more about the emotional impact someone in that situation might experience?" or "What might be going through her mind?"
  - Revised Probe for Social/Reputational Impact: "Beyond her personal feelings, how might this affect her day-to-day life or her connections with others?" or "What kind of social consequences might she face?"
  - Revised Probe for Safety Concerns: "Could there be any safety implications for the person in the video?" or "Could this situation make her feel more or less safe in any way?"
  - Revised Probe for Broader Societal Impact: "Thinking bigger picture, does this kind of thing have any wider effects on how people in general think about or treat each other online?" or "Does this raise any broader concerns for you?"
- Follow-up: "How serious do you consider this type of harm to be, and why?"
  - Where do you thing the harm begins - at generation or at distribution?
- As women, when you consider this scenario, what aspects of the harm do you think might be particularly salient or concerning to other women? + [next]
- Are there any differences in how you personally perceive the harm compared to how you think men might?"
  - Now, thinking about this same scenario – a non-consensual, sexually explicit deepfake of an adult woman being shared online – how do you think men, in general, might *perceive* the harm in this situation? Would their understanding or level of concern be the same as yours, or different? Why do you think that?
- Scenario 3 (Adult Male Target): "Now, let's switch it around. Imagine a realistic, sexually explicit video of an adult man is created using AI without his consent and shared online. What kinds of harm do you think he might experience?"
  - Probe: "What specific emotional, social, reputational, or safety issues might arise for him?" [same probes as before]
  - Considering this scenario with a male target, do you think your perception of the harm is similar to or different from how you perceived the harm to the woman in the first scenario? How might men's perceptions of the harm in this male scenario compare to your own?

- Exploring Underlying Reasons: "What do you think might explain any similarities or differences in how men and women perceive the harm caused by these types of deepfakes?" (This encourages deeper reflection on gendered perspectives and potential biases.)
- Worst Case Scenario (Gendered Comparison): "Looking ahead, do you foresee any differences in how the widespread creation of non-consensual deepfakes might impact men and women in the long run? What are your biggest concerns for each gender?

<u>Perceptions of Use and Accessibility & Impact on Minors (Approx 20-30)</u>

Incl a short trasntition/ introduction to this section - emphasis that the material is a lot more sentitive, and that all opinions and voices will be respected and heard - however, if they feel that the conversation is too much, that they are welcome to voice their concerns.

- Scenario 4 (Minor Target): "Now, imagine a similar situation, but the target of the deepfake is a minor (child or teenager). How does your perception of the harm change, if at all? What specific concerns do you have in this situation?"
    - Probe for specific vulnerabilities: "What makes a child or teenager particularly vulnerable in this kind of situation?" (Consider their developmental stage, understanding of consent, potential for long-term psychological impact, etc.)
    - Probe for legal implications: "What are your thoughts on the legal consequences for creating and sharing this kind of content when it involves a minor?"
    - Probe for long-term impact: "What kind of lasting effects, emotionally or socially, might this have on a young person?"
    - Ask again about generation vs distribution? Does their perception of this change
- *In 2024, there was a significant case in South Korea involving AI deepfakes called the 'Nth Room' scandal. Reports indicated that hundreds of thousands of sexually explicit deepfake images and videos were created and shared, often targeting women and minors. It was also reported that students from over 500 schools were involved in producing or possessing this material, and around 73% of the individuals arrested in connection with this were minors.*
    - How does hearing that statistic – 73% of those involved were minors – make you feel?
    - Does this statistic of minor involvement align with your understanding of accessibility to these technologies? Or does this information change your perception of the risks associated with deepfakes and minors in any way?
        - When we consider minors and deepfakes, does it change your perception of the harm if a young person is involved in creating versus being the victim of these? How so?
    - What does this statistic suggest to you about the accessibility and potential normalization of deepfakes among young people?

- Minors as Victims and Perpetrators: "We've seen examples like the 'Nth Room' case where minors were involved in both creating and being victims of deepfakes. What are your broader thoughts on the involvement of minors as both victims and perpetrators in deepfake-related crimes? What underlying factors do you think contribute to this dual role?"
    - Probes: (Explore ideas around digital literacy, online culture, normalization of harmful content.)
    - What might motivate a young person to create deepfakes, even harmful ones?
    - What are your thoughts on minors being involved in sharing or spreading deepfake content?
    - What unique challenges might a minor face if they become a victim of a deepfake created or shared by someone their own age?
- Responsibility and Prevention: "Who do you think bears the responsibility for addressing the issue of deepfakes, especially concerning minors? What kind of measures (educational, technological, legal) do you think could be effective in preventing harm?"
    - What role should parents, schools, tech companies, and the legal system play?
- Worst Case Scenario: Thinking about the future, what are your biggest concerns about the long-term consequences of the widespread creation and distribution of non-consensual deepfakes, specifically for how it might impact the safety and well-being of future generations of young people, both male and female?
    - "Thinking about the future, what do you see as some of the most concerning potential long-term consequences of the widespread creation and distribution of non-consensual deepfakes for women and girls?"

(Approx. 5-10 minutes) Wrap-up:

- "Is there anything else you'd like to share about your perceptions of deepfakes or their impact on women and minors that we haven't discussed?"
- Thank participants for their time and valuable contributions.
- Briefly explain the next steps of the research (data analysis).

2. Focus Group Question Guide: Male Group

Introduction & Ground Rules: (Approx. 10 minutes)

- Welcome participants and thank them for their time.
    - Hi everyone, thanks so much for taking the time to be here today. My name is { X }, and my colleague { note taker or helper } and I are conducting this discussion as part of a research project exploring how people perceive new technologies. Your honest opinions and experiences are really valuable to us."
- Briefly explain the purpose of the focus group (understanding perceptions of new technologies).

- Reiterate confidentiality and anonymity.
- Explain the format (open discussion, no right or wrong answers).
- Ask participants to briefly introduce themselves (name and study/faculty is all thats needed).
- "To start, could you briefly share your first name and maybe one thing you've heard or read about AI in the news recently?" (This can help ease participants in, but skip if time is tight).
- Understanding the Technology: "How familiar are you with how AI deepfakes are made? What's your general understanding of the technology involved?"
- Perceived Accessibility: "How easy do you think it is for the average person, including teenagers, to encounter deepfakes online? What about the ability to create them?" (Probe for awareness of user-friendly AI tools, online platforms.)

<u>(Approx. 15-20 minutes) Perceptions of Harm:</u>

- Opening Question: "When you hear the term 'AI-generated deepfake,' what are your initial thoughts or feelings?"
- Scenario 1 (Adult Female Target): "Imagine a situation where a realistic, sexually explicit video of an adult woman is created using AI without her consent and shared online. What kind of harm do you think this could cause her?" (Probe for different types of harm: emotional, psychological, social, reputational, safety, etc.)
- Follow-up: "How serious do you consider this type of harm to be, and why?"
- Scenario 2 (Minor Target): "Now, imagine a similar situation, but the target of the deepfake is a minor. How does your perception of the harm change, if at all? What specific concerns do you have in this situation?" (Probe for specific vulnerabilities of minors, legal implications, long-term impact.)
- Comparative Question: "Do you think women, in general, might experience the harm from non-consensual deepfakes differently than men? If so, in what ways?" (Encourage discussion about societal power dynamics, gendered experiences online, etc.)
- Worst Case Scenario: "Thinking about the future, what do you see as some of the most concerning potential long-term consequences of the widespread creation and distribution of non-consensual deepfakes for women and girls?"

<u>Perceptions of Harm: (Approx. 25-35 minutes)</u>

- Opening Question: "When you hear the term 'AI-generated deepfake,' what are your initial thoughts or feelings? What comes to mind?"
    - Probe: can you tell us more about why this is your initial reaction?
- *Explain a bit about what deepfake technologies to make sure everyone has a base understanding*
    - *"To be written – Just to quickly ensure we're all on the same page, 'deepfakes' are realistic-looking digital creations where someone's image or likeness has been manipulated using AI, often to create fabricated videos or images."*

- Scenario 1 (General Adult Target - Initial Reaction): "Imagine you see a realistic, sexually explicit video of an adult is created using AI without their consent being shared online. What is your inital reaction to this scenario?"
    - General Probe: "What are the first things that come to mind when you picture this happening?" (Keep intentionally Gender Neutral to see if they already assume a victim's gender"
- Scenario 2 (Adult Female Target) - if vitcim one remains gender neutral): "Imagine a situation where a realistic, sexually explicit video of an adult woman is created using AI without her consent and shared online. What kind of harm do you think this could cause her?" (Probe for different types of harm: emotional, psychological, social, reputational, safety, etc.)
    - If the initial responses are general/ if the initial victim is gendered:
    - Revised Probe for Emotional/Psychological Impact: "Could you tell us more about the emotional impact someone in that situation might experience?" or "What might be going through her mind?"
    - Revised Probe for Social/Reputational Impact: "Beyond her personal feelings, how might this affect her day-to-day life or her connections with others?" or "What kind of social consequences might she face?"
    - Revised Probe for Safety Concerns: "Could there be any safety implications for the person in the video?" or "Could this situation make her feel more or less safe in any way?"
    - Revised Probe for Broader Societal Impact: "Thinking bigger picture, does this kind of thing have any wider effects on how people in general think about or treat each other online?" or "Does this raise any broader concerns for you?"
- Follow-up: "How serious do you consider this type of harm to be, and why?"
    - Where do you thing the harm begins - at generation or at distribution?
- As men, when you consider this scenario, what aspects of the harm do you think might be particularly salient or concerning to other men? + [next]
- Are there any differences in how you personally perceive the harm compared to how you think women might perceive harm in this situation?"
    - Would their understanding or level of concern be the same as yours, or different? Why do you think that?
- Scenario 3 (Adult Male Target): "Now, let's switch it around. Imagine a realistic, sexually explicit video of an adult man is created using AI without his consent and shared online. What kinds of harm do you think he might experience?"
    - Probe: "What specific emotional, social, reputational, or safety issues might arise for him?" [same probes as before]

    - Considering this scenario with a male target, do you think your perception of the harm is similar to or different from how you perceived the harm to the woman in the first scenario? How might women's perceptions of the harm in this male scenario compare to your own?
- Exploring Underlying Reasons: "What do you think might explain any similarities or differences in how men and women perceive the harm caused by these types of deepfakes?" (This encourages deeper reflection on gendered perspectives and potential biases.)
- Worst Case Scenario (Gendered Comparison): "Looking ahead, do you foresee any differences in how the widespread creation of non-consensual deepfakes might impact men and women in the long run? What are your biggest concerns for each gender?

<u>Perceptions of Use and Accessibility & Impact on Minors (Approx 20-30)</u>

- *Incl a short trasntition/ introduction to this section - emphasis that the material is a lot more sentitive, and that all opinions and voices will be respected and heard - however, if they feel that the conversation is too much, that they are welcome to voice their concerns.*
- Scenario 4 (Minor Target): "Now, imagine a similar situation, but the target of the deepfake is a minor (child or teenager). How does your perception of the harm change, if at all? What specific concerns do you have in this situation?"
    - Probe for specific vulnerabilities: "What makes a child or teenager particularly vulnerable in this kind of situation?" (Consider their developmental stage, understanding of consent, potential for long-term psychological impact, etc.)
    - Probe for legal implications: "What are your thoughts on the legal consequences for creating and sharing this kind of content when it involves a minor?"
    - Probe for long-term impact: "What kind of lasting effects, emotionally or socially, might this have on a young person?"
    - Ask again about generation vs distribution? Does their perception of this change
- *In 2024, there was a significant case in South Korea involving AI deepfakes called the 'Nth Room' scandal. Reports indicated that hundreds of thousands of sexually explicit deepfake images and videos were created and shared, often targeting women and minors. It was also reported that students from over 500 schools were involved in producing or possessing this material, and around 73% of the individuals arrested in connection with this were minors.*
    - How does hearing that statistic – 73% of those involved were minors – make you feel?
    - Does this statistic of minor involvement align with your understanding of accessibility to these technologies? Or does this information change your perception of the risks associated with deepfakes and minors in any way?

        - When we consider minors and deepfakes, does it change your perception of the harm if a young person is involved in creating versus being the victim of these? How so?
    - What does this statistic suggest to you about the accessibility and potential normalization of deepfakes among young people?
- Minors as Victims and Perpetrators: "We've seen examples like the 'Nth Room' case where minors were involved in both creating and being victims of deepfakes. What are your broader thoughts on the involvement of minors as both victims and perpetrators in deepfake-related crimes? What underlying factors do you think contribute to this dual role?"
    - Probes: (Explore ideas around digital literacy, online culture, normalization of harmful content.)
    - What might motivate a young person to create deepfakes, even harmful ones?
    - What are your thoughts on minors being involved in sharing or spreading deepfake content?
    - What unique challenges might a minor face if they become a victim of a deepfake created or shared by someone their own age?
- Responsibility and Prevention: "Who do you think bears the responsibility for addressing the issue of deepfakes, especially concerning minors? What kind of measures (educational, technological, legal) do you think could be effective in preventing harm?"
    - What role should parents, schools, tech companies, and the legal system play?
- Worst Case Scenario: Thinking about the future, what are your biggest concerns about the long-term consequences of the widespread creation and distribution of non-consensual deepfakes, specifically for how it might impact the safety and well-being of future generations of young people, both male and female?
    - "Thinking about the future, what do you see as some of the most concerning potential long-term consequences of the widespread creation and distribution of non-consensual deepfakes for women and girls?"

<u>(Approx. 5-10 minutes) Wrap-up:</u>

- "Is there anything else you'd like to share about your perceptions of deepfakes or their impact on women and minors that we haven't discussed?"
- Thank participants for their time and valuable contributions.
- Briefly explain the next steps of the research (data analysis).

3. Focus Group Question Guide: Mixed-Gender Group (Control Group)

<u>Introduction & Ground Rules: (Approx. 10 minutes)</u>

- Welcome participants and thank them for their time.

- Hi everyone, thanks so much for taking the time to be here today. My name is { X }, and my colleague { note taker or helper } and I are conducting this discussion as part of a research project exploring how people perceive new technologies. Your honest opinions and experiences are really valuable to us."

- Briefly explain the purpose of the focus group (understanding perceptions of new technologies).
- Reiterate confidentiality and anonymity.
- Explain the format (open discussion, no right or wrong answers).
- Ask participants to briefly introduce themselves (name and study/faculty is all thats needed).
- "To start, could you briefly share your first name and maybe one thing you've heard or read about AI in the news recently?" (This can help ease participants in, but skip if time is tight).
- Understanding the Technology: "How familiar are you with how AI deepfakes are made? What's your general understanding of the technology involved?"
- Perceived Accessibility: "How easy do you think it is for the average person, including teenagers, to encounter deepfakes online? What about the ability to create them?" (Probe for awareness of user-friendly AI tools, online platforms.)

<u>Perceptions of Harm: (Approx. 25-35 minutes)</u>

- Opening Question: "When you hear the term 'AI-generated deepfake,' what are your initial thoughts or feelings? What comes to mind?"
    - Probe: can you tell us more about why this is your initial reaction?
- *Explain a bit about what deepfake technologies to make sure everyone has a base understanding*
    - *"To be written – Just to quickly ensure we're all on the same page, 'deepfakes' are realistic-looking digital creations where someone's image or likeness has been manipulated using AI, often to create fabricated videos or images."*
- Scenario 1 (General Adult Target - Initial Reaction): "Imagine you see a realistic, sexually explicit video of an adult is created using AI without their consent being shared online. What is your inital reaction to this scenario?"
    - General Probe: "What are the first things that come to mind when you picture this happening?" (Keep intentionally Gender Neutral to see if they already assume a victim's gender"
- Scenario 2 (Adult Female Target) - if vitcim one remains gender neutral): "Imagine a situation where a realistic, sexually explicit video of an adult woman is created using AI without her consent and shared online. What kind of harm do you think this could cause her?" (Probe for different types of harm: emotional, psychological, social, reputational, safety, etc.)
    - If the initial responses are general/ if the initial victim is gendered:

    - Revised Probe for Emotional/Psychological Impact: "Could you tell us more about the emotional impact someone in that situation might experience?" or "What might be going through her mind?"
    - Revised Probe for Social/Reputational Impact: "Beyond her personal feelings, how might this affect her day-to-day life or her connections with others?" or "What kind of social consequences might she face?"
    - Revised Probe for Safety Concerns: "Could there be any safety implications for the person in the video?" or "Could this situation make her feel more or less safe in any way?"
    - Revised Probe for Broader Societal Impact: "Thinking bigger picture, does this kind of thing have any wider effects on how people in general think about or treat each other online?" or "Does this raise any broader concerns for you?"
- Follow-up: "How serious do you consider this type of harm to be, and why?"
    - Where do you thing the harm begins - at generation or at distribution?
- Considering this scenario of a non-consensual deepfake of an adult woman, do you think men and women in this group might have similar or different perspectives on the types and severity of harm? Let's hear some different viewpoints.
- Scenario 3 (Adult Male Target): "Now, let's switch it around. Imagine a realistic, sexually explicit video of an adult man is created using AI without his consent and shared online. What kinds of harm do you think he might experience?"
    - Probe: "What specific emotional, social, reputational, or safety issues might arise for him?" [same probes as before]
    - Considering this scenario with a male target, do you think your perception of the harm is similar to or different from how you perceived the harm to the woman in the first scenario?
    - [following] When we compare the potential harm to a man in this scenario versus the potential harm to a woman in the previous one, do you notice any similarities or differences in your own perceptions or in what you've heard from others in the group?
- Exploring Underlying Reasons: "What do you think might explain any similarities or differences in how men and women perceive the harm caused by these types of deepfakes?" (This encourages deeper reflection on gendered perspectives and potential biases.)
- Worst Case Scenario (Gendered Comparison): "Looking ahead, do you foresee any differences in how the widespread creation of non-consensual deepfakes might impact men and women in the long run? What are your biggest concerns for each gender?

<u>Perceptions of Use and Accessibility & Impact on Minors (Approx 20-30)</u>

- *Incl a short trasntition/ introduction to this section - emphasis that the material is a lot more sentitive, and that all opinions and voices will be respected and heard - however, if they feel that the conversation is too much, that they are welcome to voice their concerns.*
- Scenario 4 (Minor Target): "Now, imagine a similar situation, but the target of the deepfake is a minor (child or teenager). How does your perception of the harm change, if at all? What specific concerns do you have in this situation?"
    - Probe for specific vulnerabilities: "What makes a child or teenager particularly vulnerable in this kind of situation?" (Consider their developmental stage, understanding of consent, potential for long-term psychological impact, etc.)
    - Probe for legal implications: "What are your thoughts on the legal consequences for creating and sharing this kind of content when it involves a minor?"
    - Probe for long-term impact: "What kind of lasting effects, emotionally or socially, might this have on a young person?"
    - Ask again about generation vs distribution? Does their perception of this change
- *In 2024, there was a significant case in South Korea involving AI deepfakes called the 'Nth Room' scandal. Reports indicated that hundreds of thousands of sexually explicit deepfake images and videos were created and shared, often targeting women and minors. It was also reported that students from over 500 schools were involved in producing or possessing this material, and around 73% of the individuals arrested in connection with this were minors.*
    - How does hearing that statistic – 73% of those involved were minors – make you feel?
    - Does this statistic of minor involvement align with your understanding of accessibility to these technologies? Or does this information change your perception of the risks associated with deepfakes and minors in any way?
        - When we consider minors and deepfakes, does it change your perception of the harm if a young person is involved in creating versus being the victim of these? How so?
    - What does this statistic suggest to you about the accessibility and potential normalization of deepfakes among young people?
- Minors as Victims and Perpetrators: "We've seen examples like the 'Nth Room' case where minors were involved in both creating and being victims of deepfakes. What are your broader thoughts on the involvement of minors as both victims and perpetrators in deepfake-related crimes? What underlying factors do you think contribute to this dual role?"
    - Probes: (Explore ideas around digital literacy, online culture, normalization of harmful content.)
    - What might motivate a young person to create deepfakes, even harmful ones?

    - What are your thoughts on minors being involved in sharing or spreading deepfake content?
    - What unique challenges might a minor face if they become a victim of a deepfake created or shared by someone their own age?
- Responsibility and Prevention: "Who do you think bears the responsibility for addressing the issue of deepfakes, especially concerning minors? What kind of measures (educational, technological, legal) do you think could be effective in preventing harm?"
    - What role should parents, schools, tech companies, and the legal system play?
- Worst Case Scenario: Thinking about the future, what are your biggest concerns about the long-term consequences of the widespread creation and distribution of non-consensual deepfakes, specifically for how it might impact the safety and well-being of future generations of young people, both male and female?
    - "Thinking about the future, what do you see as some of the most concerning potential long-term consequences of the widespread creation and distribution of non-consensual deepfakes for women and girls?"

<u>(Approx. 5-10 minutes) Wrap-up:</u>

- "Is there anything else you'd like to share about your perceptions of deepfakes or their impact on women and minors that we haven't discussed?"
- Thank participants for their time and valuable contributions.
- Briefly explain the next steps of the research (data analysis).

14. Appendix F – Critical Discourse Articles

https://www.nbcnews.com/news/us-news/little-recourse-teens-girls-victimized-ai-deepfake-nudes-rcna126399 - NBC (Source 1)

For teen girls victimized by 'deepfake' nude photos, there are few, if any, pathways to recourse in most states

By Melissa Chan and Kat Tenbarge

Nov. 23, 2023 – 852 words

Teenage girls in the U.S. who are increasingly being targeted or threatened with fake nude photos created with artificial intelligence or other tools have limited ways to seek accountability or recourse, as schools and state legislatures struggle to catch up to the new technologies, according to legislators, legal experts and one victim who is now advocating for a federal bill.

Since the 2023 school year kicked into session, cases involving teen girls victimized by the fake nude photos, also known as deepfakes, have proliferated worldwide, including at high schools in New Jersey and Washington state.

Local police departments are investigating the incidents, lawmakers are racing to enact new measures that would enforce punishments against the photos' creators, and affected families are pushing for answers and solutions.

Unrealistic deepfakes can be made with simple photo-editing tools that have existed for years. But two school districts told NBC News that they believe fake photos of teens that have affected their students were AI-generated.

AI technology is becoming more widely available, such as stable diffusion (open-source technology that can produce images from text prompts) and "face-swap" tools that can put a victim's face in place of a pornographic performer's face in a video or photo.

Apps that purport to "undress" clothed photos have also been identified as possible tools used in some cases and have been found available for free on app stores. These modern deepfakes can be more realistic-looking and harder to immediately identify as fake.

"I didn't know how complex and scary AI technology is," said Francesca Mani, 15, a sophomore at New Jersey's Westfield High School, where more than 30 girls learned on Oct. 20 that they may have been depicted in explicit, AI-manipulated images.

"I was shocked because me and the other girls were betrayed by our classmates," she said, "which means it could happen to anyone by anyone."

Politicians and legal experts say there are few, if any, pathways to recourse for victims of AI-generated and deepfake pornography, which often attaches a victim's face to a naked body.

The photos and videos can be surprisingly realistic, and according to Mary Anne Franks, a legal expert in nonconsensual sexually explicit media, the technology to make them has become more sophisticated and accessible.

A month after the incident at Westfield High School, Francesca and her mother, Dorota Mani, said they still do not know the identities or the number of people who created the images, how many were made, or if they still exist. It's also unclear what punishment the school district doled out, if any.

The Town of Westfield directed comment to Westfield Public Schools, which declined to comment. Citing confidentiality, the school district previously told NBC New York that it "would not release any information about the students accused of creating the fake nude photos, or what discipline they are facing."

Superintendent Raymond Gonzalez told the news outlet that the district would "continue to strengthen our efforts by educating our students and establishing clear guidelines to ensure that these new technologies are used responsibly in our schools and beyond."

In an email obtained by NBC News, Mary Asfendis, the high school's principal, told parents on Oct. 20 that it was investigating claims by students that some of their peers had used AI to create pornographic images from original photos.

At the time, school officials believed any created images had been deleted and were not being circulated, according to the memo.

"This is a very serious incident," Asfendis wrote, as she urged parents to discuss their use of technology with their children. "New technologies have made it possible to falsify images and students need to know the impact and damage those actions can cause to others."

While Francesca has not seen the image of herself or others, her mother said she was told by Westfield's principal that four people identified Francesca as a victim. Francesca has filed a police report, but neither the Westfield Police Department nor the prosecutor's office responded to requests for comment.

New Jersey State Sen. Jon Bramnick said law enforcement expressed concerns to him that the incident would only rise to a "cyber-type harassment claim, even though it really should reach the level of a more serious crime."

"If you attach a nude body to a child's face, that to me is child pornography," he said.

The Republican lawmaker said state laws currently fall short of punishing the content creators, even though the damage inflicted by real or manipulated images can be the same.

"It victimizes them the same way people who deal in child pornography do. It's not only offensive to the young person, it defames the person. And you never know what's going to happen to that photograph," he said. "You don't know where that is once it's transmitted, when it's going to come back and haunt the young girl."

A pending state bill in New Jersey, Bramnick said, would ban deepfake pornography and impose criminal and civil penalties for nonconsensual disclosure. Under the bill, a person convicted of the crime would face three to five years in jail and/or a $15,000 fine, he said.

https://www.koreaherald.com/article/3832399 - Korean Herald (Source 2)

4 middle schoolers booked for producing deepfake porn of classmates

By Lee Jung-joo

24 October 2024 – 562 words

More than 1,700 digital sex offenses at schools reported since 2021.

Four middle school students in Namyangju, Gyeonggi Province, have been booked by the police for producing, possessing and distributing deepfake pornography, according to Gyeonggi Bukbu Provincial Police Agency.

Police officials confirmed on Saturday that two of the four individuals are suspected of using photos of their female classmates to create sexually explicit deepfake content since November 2023.

The two students are additionally charged with possession of the deepfakes and sharing them with the other two students, who themselves are charged with possession of the content.

The four middle school students have been booked for violating the Act on Special Cases Concerning the Punishment of Sexual Crimes. Under South Korean law, it is a crime to possess sexually explicit deepfake images of minors.

Based on an investigation, the police have so far identified nine victims who are all female middle school students. However, the police added that the number could increase as the investigation continues.

Regarding the investigation, police officials added that detectives had searched the suspects' homes and their mobile phones for evidence of further offenses.

The case is also being investigated by the Gyeonggi Bukbu Provincial Police Agency's Cyber Investigation Bureau after initially being reported to the Namyangju Bukbu Police Station.

The incident came to light in August when the victims reported to the school that four students had created sexually explicit deepfake images of them and their friends.

The school reported the case to the Guri Namyangju Office of Education, which formed a School Violence Countermeasures Review Committee. The committee decided to transfer the two students who created the deepfake content and suspend the two others accused of possessing the content.

However, local media reports said the victims' parents have filed a complaint, claiming their children suffered "secondary victimization" due to the school's delayed response -- such as taking two months to separate the female victims from the suspects. Secondary victimization refers to further trauma experienced by victims due to insensitivity, blaming or dismissive attitudes from others.

From 2021 to August 2024, 1,727 digital sex offenses by students were reported to the School Violence Countermeasures Review Committee, according to data provided by 16 provincial and metropolitan education offices to Rep. Kang Kyung-sook of the Rebuilding Korea Party.

Out of the total number of digital sex offenses, in 765 cases, or 44.3 percent, the committee required "severe punishment" to be taken against the perpetrators.

In South Korea, cases of violence at schools are reviewed by a School Violence Countermeasures Review Committee, which meets after an investigation to confirm the incident and decide on disciplinary actions.

Punishments range from a written apology, no-contact orders or community service to more severe measures such as suspension, class transfer, school transfer or expulsion.

According to Kang, some of the reported digital sex offenses included the creation and distribution of sexually explicit deepfake videos and using them to threaten the victim, and the illegal creation and distribution of deepfake and illegally filmed content online. There were also a few cases where the perpetrators sent messages constituting sexual harassment.

“Not only physical violence but also digital sex offenses committed online also constitute a type of school violence,” said Kang. “Active attention as well as education for the students must be provided by education authorities to prevent further victimization of students and to ultimately prevent school violence.”

https://www.cnn.com/2023/11/04/us/new-jersey-high-school-deepfake-porn/index.html - CNN

(Source 3)

High schooler calls for AI regulations after manipulated pornographic images of her and others shared online

By Skylar Harris and Artemis Moshtaghian

November 4th 2023 – 573 words

A student at a New Jersey high school is calling for federal legislation to address AI generated pornographic images after she says photos of her and other female classmates were manipulated and possibly shared online over the summer.

Westfield High School student Francesca Mani, 14, and her mother, Dorota, have expressed frustration over what they say is a lack of legal recourse in place to protect victims of AI-generated pornography.

“In this situation, there was some boys or a boy — that’s to be determined — who created, without the consent of the girls, inappropriate images,” Dorota said, speaking with CNN’s Michael Smerconish Saturday.

Francesca, who said she was among more than 30 female students at Westfield High School whose photos were manipulated and possibly shared publicly, is demanding accountability from the school and local, state, and government officials.

School administrators initially became aware of the incident on October 20 when students informed them the images were created and possibly shared over the summer.

“There was a great deal of concern about who had images created of them and if they were shared,” Westfield Principal Mary Asfendis wrote in a letter to students and parents sent on October 20. “At

this time, we believe that any created images have been deleted and are not being circulated. This is a very serious incident."

Westfield High School has since conducted its own investigation and the Westfield Police Department and the school's appointed resource officer "were immediately notified and consulted throughout the investigation," according to school spokesperson Mary Ann McGann. CNN has contacted the Westfield Police Department for comment.

McGann told CNN the school is not able "to provide specific details on the number of students involved and any disciplinary actions imposed, as matters involving students are confidential."

The school provided CNN with a statement from Superintendent Dr. Raymond González, who said, "All school districts are grappling with the challenges and impact of artificial intelligence and other technology available to students at any time and anywhere."

González added, "The Westfield Public School District has safeguards in place to prevent this from happening on our network and school-issued devices. We continue to strengthen our efforts by educating our students and establishing clear guidelines to ensure that these new technologies are used responsibly in our schools and beyond."

Dorota said she's proud of her daughter for speaking up and advocating not only for herself, but on behalf of other young girls who have also been victimized by AI generated deepfake pornographic content.

"I think this issue is more complex than just Westfield High School, and this is our time and opportunity to treat it as a teachable platform, to shed the light on this important issue," she said.

Dorota said her daughter has urged her to see if there are any laws in New Jersey protecting against deepfake images or videos and has also written a letter to President Joe Biden asking him to urge state governors to make sure there are laws in place to protect underage girls and boys.

CNN has reached out to the New Jersey Union County Prosecutor's Office and the White House for comment.

Intelligence officials in the US have warned about the sharp rise in deepfake videos, which may look convincingly real but are generated using artificial intelligence. In California, bills have been written to combat the use of deepfakes in nonconsensual pornography.

https://www.cnn.com/2024/06/13/australia/australia-boy-arrested-deepfakes-schoolgirls-intl-hnk/index.html - CNN (Source 4)

Teenager questioned after explicit AI deepfakes of dozens of schoolgirls shared online

By Angus Watson and Hilary Whiteman

June 13 2024 – 690 words

Australian authorities are investigating the distribution of deepfake pornographic images of around 50 schoolgirls, allegedly created by a teenager using artificial intelligence.

The discovery comes as the federal government pushes for new laws to impose prison sentences on offenders who create and share images made by AI tools to humiliate and denigrate victims.
Other countries, including the United States, are attempting to address an alarming rise in deepfake porn, where nude deepfakes of schoolgirls have been created and shared – in some instances, allegedly by schoolboys.
Victoria Police confirmed they had arrested and released a teenager "in relation to explicit images being circulated online" pending further inquiries.
The images were reportedly created using photos posted to social media of 50 female students of Bacchus Marsh Grammar, a co-educational school on the outskirts of Melbourne in Victoria.
The school's principal Andrew Neal told the Australian Broadcasting Corporation (ABC) that the victims were girls in grades 9 to 12, indicating a possible age range of between 14 and 18.
The boy's age and identity are unknown, but Neal told the ABC that "logic would suggest that the [offender] is someone at the school."
Speaking to the ABC on Wednesday, the mother of a 16-year-old female Bacchus Marsh Grammar student, whose image wasn't used, said her daughter vomited when she saw the "mutilated" pictures online.
"I went and picked my daughter up from a sleepover and she was very upset, and she was throwing up and it was incredibly graphic," the mother told ABC Radio Melbourne, giving only her first name, Emily.
The school said in a statement that it was offering counseling to the students and assisting police with their investigation.
"The wellbeing of Bacchus Marsh Grammar students and their families is of paramount importance to the school and is being addressed," the statement said.
Legal fight to stop deepfakes
Social media companies, including X and Meta say all nonconsensual pornography is banned on their platforms, but explicit AI-generated images continue to quickly spread online
Last November, New Jersey high school student Francesca Mani, 14, led public demands for a federal crackdown in the US against AI-generated deepfake pornography, saying images of her and dozens of her classmates at Westfield High School had been manipulated.
High-profile victims of explicit doctored images include Taylor Swift and New York Congresswoman Alexandria Ocasio-Cortez.
In March, Ocasio-Cortez introduced federal legislation – the Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024 (DEFIANCE Act) – to give victims the power to sue people who create non-consensual deepfakes of them.
However, the bi-partisan legislation, backed by senior Republicans, failed to pass a motion for unanimous consent Wednesday, according to a statement by the Senate Committee on the Judiciary.
Victoria is the only Australian state where sharing deepfake pornography is a criminal offense.

In 2022, the state government introduced three-year jail terms for using technology to generate or share child abuse material, or sexually explicit material without consent.

This month, the Australian government introduced legislation to criminalize the distribution of deepfake pornography nationwide.

Under the proposed law, offenders could face up to six years in prison for sharing non-consensual sexually explicit deepfake material.

If the offender has also created the deepfake content that is shared without permission, the sentence could rise to seven years in prison.

It's part of the country's response to gender-based violence that Prime Minister Anthony Albanese has called a "national crisis."

So far this year, 35 women have been killed, according to the Counting Dead Women project – many allegedly by current or former partners.

Just last month, the state government appointed a Parliamentary Secretary for Men's Behavior Change, in an Australian first.

On his appointment, MP Tim Richardson said he would focus on the impact of the internet and social media on men's attitudes toward women.

In a statement on Wednesday, Victoria State Premier Jacinta Allan said the alleged actions of the teenager were "disgraceful and misogynistic."

"Women and girls deserve respect in class, online and everywhere else in our community, which is why we have made laws against this behavior and we are teaching respectful relationships in schools to stop violence before it starts," Allan said.

https://www.theguardian.com/australia-news/article/2024/jun/12/schoolboy-arrested-after-allegedly-posting-fake-explicit-images-of-female-students-ntwnfb - The Guardian (Source 5)

Bacchus Marsh Grammar: schoolboy arrested after 50 female students allegedly targeted in fake explicit AI photos scandal

By Jordyn Beazley and Rafqa Touma

Wed 12 Jun 2024 – 723 words

A teenage boy has been arrested and then released after fake explicit images, described as "mutilated" and "incredibly graphic", were allegedly circulated on social media using the likenesses of about 50 female students from a private school in regional Victoria.

The principal of Bacchus Marsh Grammar, Andrew Neal, told the ABC about 50 girls had been targeted.

The nude images appeared to have been created using artificial intelligence and photos of the girls' faces taken from social media sites, and were then circulated online.

“[The girls] should be able to learn and go about their business without this kind of nonsense,” Neal told the ABC.

A woman named Emily, the parent of one Bacchus Marsh Grammar student and a trauma therapist, said she saw the photos when she picked up her 16-year-old daughter from a sleepover on Saturday night.

Emily had a bucket in the car for her daughter who was “sick to her stomach” on the drive home, she told ABC Radio Melbourne on Wednesday.

“She was very upset, and she was throwing up. It was incredibly graphic.”

Emily’s immediate reaction was also “to be sick”.

“I mean they are children … The photos were mutilated, and so graphic. I almost threw up when I saw it.”

“Fifty girls is a lot. It is really disturbing.”

The victims’ Instagram accounts were private, Emily said. Though her daughter did not appear in the deepfake images, “there’s just that feeling of ... will this happen again? It’s very traumatising.”

“How can we reassure them that once measures are in place, it won’t happen again?”

Acting principal Kevin Richardson said in a statement: “Bacchus Marsh Grammar is taking this matter very seriously and has contacted Victoria police.

“The wellbeing of Bacchus Marsh Grammar students and their families is of paramount importance to the school and is being addressed. All students affected are being offered support from our wellbeing staff.”

Richardson said the school had not been contacted by police regarding anyone arrested in relation to the matter.

Victoria police said officers were informed that a number of images were sent to a person in the Melton area, which is about a 15-minute drive from Bacchus Marsh, via an online platform on Friday 7 June.

Police said the teenage boy was arrested in relation to the incident, but had been released pending further inquiries.

“The investigation remains ongoing,” police said.

It comes after a student from Salesian College, a Catholic boys’ school in Chadstone in Melbourne, was expelled after he used artificial intelligence to produce explicit images of a female teacher.

In June the federal government introduced legislation to ban the creation and sharing of deepfake pornography as part of measures to combat violence against women.

But the Nationals senator Matt Canavan said there was a broader cultural problem that needed to be addressed.

“It is a cultural issue across our society that, for whatever reason, the standards of behaviour are not being taught to young boys,” he told Nine on Wednesday morning.

“I wish I had the answers – I don’t – but I don’t necessarily think it’s something a government or a law can change.

“We’ve all got to chip in to try and make sure that young boys understand what it means to grow up to be a man and live by the standards that society expects.”

Technology had “supercharged” boys’ bad behaviour, Canavan said.

Victoria’s health minister, Mary-Anne Thomas, said she was “deeply distressed” to read of the incident.

“This type of behaviour is absolutely abhorrent. Schools should be safe places for children and young people to learn,” she said on Wednesday.

While Thomas said she did not want to comment specifically on the incident, she said she hoped it would be a “wake up call” to the dangers of AI-generated sexually explicit images and social media.

“We need to be having direct and real conversations with young people about respect,” she said.

“Because quite clearly, young people are accessing material on the internet, and through social media, that is influencing their behaviours in ways that I think we all agree are out of step with community expectations – and indeed are beyond that, [and are] actually causing real harm to other young people.”

Image-based abuse, including deepfakes, can be reported to eSafety, which claims “a 90% success rate in getting this distressing material down”.

https://www.koreaherald.com/article/3479503 - Korean Herald (Source 6)

School sexual assaults at record high with ‘deepfakes,’ cyberbullying

By Choi Jeong-yoon

September 25 2024 – 531 words

The number of elementary, middle and high school students reporting being bullied at school increased for the fourth straight year, pushing the corresponding rate to 2 percent for the first time in 11 years.

With 1 out of 50 children having experienced school violence, the proportion of victim students was higher among younger students. The modality of violence became more insidious and adroit as the ratio of verbal and cyber violence took over physical bullying, according to the report by the Ministry of Education on Wednesday.

The ministry surveyed 3.98 million students nationwide from the fourth grade in elementary school to fourth-year students in high school in the first half of this year.

In 2024, the prevalence of school bullying in the national survey was 2.1 percent, up 0.2 percentage points from last year. The rate of students answering they experienced violence dropped from 2.2 percent in 2013 to 0.9 percent in 2016 and 2017 before rising again to 1.6 percent in 2019. The

number again fell to 0.9 percent in 2020 due to an increase in virtual learning from the COVID-19 pandemic.

However, as schools returned to in-person learning, the rate rose for four consecutive years to 1.1 percent in 2021, 1.7 percent in 2022, 1.9 percent in 2023, and 2.1 percent in 2024.

The rates of victimization of peer violence increased as students got younger, with elementary school reporting 4.2 percent, middle school 1.6 percent, and high school at 0.5 percent, up 0.3 percentage points, 0.3 percentage points, and 0.1 percentage points, respectively, from last year.

Among the types of victimization, verbal abuse accounted for the largest share, taking up 39.4 percent of the total. This is up 2.3 percentage points from last year.

Amid the recent controversy over deepfake sexual exploitation, cyberbullying, especially among high school students, also increased. Putting it in third place of all violations, cyberbullying accounted for 7.4 percent of the total.

Deepfakes, which have recently become controversial in the country due to their widespread distribution through Telegram group chat rooms, also fall under the category of cyberbullying.

In particular, the victimization rate of cyberbullying was higher among high school students at 10.4 percent, compared to elementary school students at 6.3 percent and middle school students at 9.2 percent.

By type of cyberbullying, the most common types were "cyber verbal abuse" at 38.1 percent, "cyber defamation" at 16.6 percent, and "cyber ostracism" at 16.1 percent.

In the case of verbal violence and sexual violence, the ministry attributed the increased sensitivity of students rather than the increase in actual incidents to the rise in numbers, as what used to be overlooked, such as sexual jokes, are now being recognized as school violence.

However, actual reports of school violence have increased as there were 61,445 reports of school violence in elementary, middle and high schools last year, a 6 percent increase from the 2022 school year, which recorded a total of 57,981 report cases, according to the ministry's other data on the status of receiving and handling school violence.

Of these, 23,579 cases were reviewed by the school committee after failing to be resolved by the school principal, an increase of 9.3 percent from the previous year.

https://www.independent.co.uk/news/world/americas/elliston-berry-deepfakes-social-media-b2566806.html - The Independent (Source 7)

Girl, 15, calls for criminal penalties after classmate made deepfake nudes of her and posted on social media

By Katie Hawkinson

Tuesday 25 June 2024 – 603 words

Last October, 14-year-old Elliston Berry woke up to a nightmare.

The teen's phone was flooded with calls and texts telling her that someone had shared fake nude images of her on Snapchat and other social media platforms.

"I was told it went around the whole school," Berry, from Texas, told Fox News. "And it was just so scary going through classes and attending school, because just the fear of everyone seeing these images, it created so much anxiety."

Last October, 14-year-old Elliston Berry woke up to a nightmare.

The teen's phone was flooded with calls and texts telling her that someone had shared fake nude images of her on Snapchat and other social media platforms.

"I was told it went around the whole school," Berry, from Texas, told Fox News. "And it was just so scary going through classes and attending school, because just the fear of everyone seeing these images, it created so much anxiety."

The teen told the outlet that after discovering what had happened, she immediately went to her parents. Her mother, Anna McAdams, told Fox News she knew the images were fake. McAdams then reached out to Snapchat several times over an eight-month period to have the photos removed.

While the deepfakes of Berry were eventually taken down, McAdams told CNN, the classmate who distributed them is facing few repercussions.

"This kid who is not getting any kind of real consequence other than a little bit of probation, and then when he's 18, his record will be expunged, and he'll go on with life, and no one will ever really know what happened," McAdams told CNN.

This week, Republican Senator Ted Cruz, Democratic Senator Amy Klobuchar and several colleagues co-sponsored a bill that would require social media companies to take down deep-fake pornography within two days of getting a report.

The Take It Down Act would also make it a felony to distribute these images, Cruz told Fox News. Perpetrators who target adults could face up to two years in prison, while those who target children could face three years.

Cruz said that what happened to Berry "is a sick and twisted pattern that is getting more and more common."

"[The bill] puts a legal obligation on the big tech companies to take it down, to remove the images when the victim or the victim's family asks for it," Cruz said. "Elliston's Mom went to Snapchat over and over and over again, and Snapchat just said, 'Go jump in a lake.' They just ignored them for eight months."

A spokesperson for Snap Inc, Snapchat's parent company, said the platform does not allow pornography and has policies that prohibit deepfakes and bullying.

The mom and daughter say legislation is essential to protecting future victims, and could have meant more serious consequences for the classmate who shared the deep-fakes.

15. Appendix G - Declaration Page: Use of Generative AI Tools in Thesis

Student Information
Name: Anja Ellwood
Student ID: 740726
Course Name: Master Thesis CM5000
Supervisor Name: João Gonçalves, PhD
Date:

Declaration:

Acknowledgment of Generative AI Tools
I acknowledge that I am aware of the existence and functionality of generative artificial intelligence (AI) tools, which are capable of producing content such as text, images, and other creative works autonomously.

GenAI use would include, but not limited to:

- Generated content (e.g., ChatGPT, Quillbot) limited strictly to content that is not assessed (e.g., thesis title).
- ~~Writing improvements, including~~ grammar and spelling corrections (e.g., Grammarly)
- Language translation (e.g., DeepL), without generative AI alterations/improvements.
- Research task assistance (e.g., finding survey scales, qualitative coding verification, debugging code)
- Using GenAI as a search engine tool to find academic articles or books (e.g.,

☐ I declare that I have used generative AI tools, specifically [Name of the AI Tool(s) or Framework(s) Used], in the process of creating parts or components of my thesis. The purpose of using these tools was to aid in generating content or assisting with specific aspects of thesis work.

☒ I declare that I have NOT used any generative AI tools and that the assignment concerned is my original work.

Signature: Anja Ellwood

Date of Signature: 30.06.2025

Extent of AI Usage

☐ I confirm that while I utilized generative AI tools to aid in content creation, the majority of the intellectual effort, creative input, and decision-making involved in completing the thesis were undertaken by me. I have enclosed the prompts/logging of the GenAI tool use in an appendix.

Ethical and Academic Integrity

☐ I understand the ethical implications and academic integrity concerns related to the use of AI tools in

coursework. I assure that the AI-generated content was used responsibly, and any content derived from these tools has been appropriately cited and attributed according to the guidelines provided by the instructor and the course. I have taken necessary steps to distinguish between my original work and the AI-generated contributions. Any direct quotations, paraphrased content, or other forms of AI-generated material have been properly referenced in accordance with academic conventions.

By signing this declaration, I affirm that this declaration is accurate and truthful. I take full responsibility for the integrity of my assignment and am prepared to discuss and explain the role of generative AI tools in my creative process if required by the instructor or the Examination Board. I further affirm that I have used generative AI tools in accordance with ethical standards and academic integrity expectations.

Signature: [digital signature]

Date of Signature: [Date of Submission]