Grokking Misogyny: Is New Zealand Law Ready for Sexualised Deepfakes?

Cassandra Mudgway Senior Lecturer, University of Canterbury*
Christopher McGavin Visiting Scholar, Victoria University of Wellington*
Andrew Lensen Senior Lecturer in Artificial Intelligence, Victoria University of Wellington*

This article examines whether New Zealand's legal framework is equipped to address the rise of nonconsensual sexualised deepfakes generated generative AI tools integrated into social media platforms. While platforms such as X and Meta increasingly deploy AI systems capable of producing realistic synthetic images, these technologies have intensified gendered abuse, disproportionately targeting women. Using a Grok-generated deepfake incident as a case study, the article evaluates the applicability and limitations of the Harmful Digital Communications Act 2015, privacy torts, defamation, and negligence. Taking a feminist legal and technofeminist lens, this article argues that existing laws have limited meaningful application to harm enabled by AI. Although the proposed Deepfake Digital Harm and Exploitation Bill may close definitional gaps, criminalisation alone cannot address systemic risks created by platform design, business incentives, and inadequate safeguards.

I Introduction

New Zealand has maintained a "light-touch" regulatory approach to Artificial Intelligence (AI), even as Generative AI (GenAI) technology continues to mature and presents new sociotechnical issues. This approach, enshrined in the country's 2025 AI Strategy,[1] is designed to encourage new economic activity and avoid the introduction of regulation that becomes quickly out-of-date as GenAI continues to improve.

One such sociotechnical issue is the increasing volume of "deepfakes", namely, GenAI-created images or videos that look real, but are not. Advances in GenAI have made the creation of realistic deepfakes widely accessible. Now, any user can log into a social media platform such as Facebook

[1] *New Zealand's Strategy for Artificial Intelligence: Investing with confidence* (Ministry of Business, Innovation and Employment, July 2025).

or X and ask a chatbot to create one in seconds. This democratisation of the technology has led to vast amounts of the internet becoming AI-generated i.e. "fake".

Recently, there have been increased reports of sexualised deepfakes on social media (e.g. platforms hosting user content and messaging services), where GenAI is used to "undress" users into underwear or lingerie. These deepfakes, predominantly of women, can be created to embarrass, degrade, or otherwise cause harm to the targeted user. The gravity of these harms and evidence of their proliferation is at odds with the "light-touch" regulatory approach that is currently favoured. To date, there has been little work done to interrogate the efficacy of laws and New Zealand's ability to respond to the weaponisation of deepfakes. It is unclear if responses available in either criminal or civil law would apply, and to what extent they would prevent harm or provide redress. This paper will canvas New Zealand legal framework to establish its suitability in our new epoch of AI-enabled harm.[2]

Our analysis is informed by feminist legal theory and technofeminism, which emphasise that laws and technologies presented as objective are embedded within, and shaped by, unequal social contexts.[3] From this perspective, harms associated with AI-driven systems are understood as foreseeable consequences of their design, training, and deployment within patriarchal social and economic structures.

In terms of structure, Part II examines how GenAI integration into social media platforms reinforces online misogyny and introduces this paper's deepfake case study: the "Grok Scenario". Part III uses the Grok Scenario to assess individual liability under New Zealand criminal and civil law, while Part IV considers platform responsibility for AI-enabled harms. The paper concludes by reflecting on the limits of New Zealand's largely reactive legal framework and the need for law reform.

[2] The focus of this paper is on adults as targets of sexualised deepfakes, so will not include discussion of child sexual abuse material. Further, the type of sexualised deepfakes discussed in this paper would likely fall short of 'objectionable' under the Films, Videos, and Publications Classification Act 1993 because they typically do not concern sexual violence or sexual conduct. Thus, this article will not discuss this Act.

[3] See for example: Catharine MacKinnon *Toward a Feminist Theory of the State* (Harvard University Press, Cambridge, 1989); Nicola Lacey *Unspeakable Subjects: Feminist Essays in Legal and Social Theory* (Hart, Oxford, 1998); Carol Smart *Feminism and the Power of Law* (Routledge, London, 1986); Judy Wajcman *Technofeminism* (John Wiley and Sons, New Jersey, 2004); Jude Brown and others (eds) *Feminist AI* (Oxford University Press, Oxford, 2023).

The legal frameworks examined are those most directly engaged by the conduct in the case study and by the question of responsibility for AI-enabled gendered harm. The Harmful Digital Communications Act 2015 (HDCA) provides the principal statutory framework for online harm and the sharing of nonconsensual intimate images, while privacy, defamation and negligence provide relevant civil avenues for addressing interference with sexual privacy, reputation and platform conduct. Examining these together allows the analysis to consider both individual redress and the extent to which existing law can respond to harms enabled by platform design.

II Social Media, Generative AI and Gendered Harm

Since the early 2010s, social media platforms have relied on AI systems to curate user feeds and moderate content.[4] Feminist scholars have long criticised these technologies for reproducing gender stereotypes, amplifying misogynistic content, and disproportionately censoring feminist, LGBTQIA+, and other marginalised forms of expression.[5] The rapid integration of GenAI into social media platforms emerged against this backdrop of existing gendered harms.

By the mid-2020s, investor enthusiasm for AI had reached fever pitch. With global AI spending projected to research nearly $2.7 trillion in 2026,[6] platforms are facing strong shareholder pressure to be seen as AI leaders. In 2024, Meta introduced Meta AI across Facebook, Instagram, and WhatsApp, while X integrated its own chatbot, Grok.[7] At the same time, increasingly sophisticated image-generation tools became available within social media ecosystems, though their widespread

[4] See for example, Trefis Team "Facebook: The Year 2016 In Review" (16 December 2016) *Forbes* <https://www.forbes.com>; Jeffrey Dunn "Introducing FBLearner Flow: Facebook's AI backbone" (9 May 2016) *Engineering at Meta* <https://engineering.fb.com>; Biao Gao and others "Artificial Intelligence in Advertising: Advancements, Challenges, and Ethical Considerations in Targeting, Personalization, Content Creation, and Ad Optimization" (2023) *SAGE Open* 1.

[5] See for example, Rena Bivens and Oliver Haimson "Baking Gender into Social Media Design: How Platforms Shape Categories for Users and Advertisers" (2016) 2 *Social Media + Society* 1; Ruth Lewis, Michael Rowe and Clare Wiper "Online Abuse of Feminists as An Emerging form of Violence Against Women and Girls" (2017) 57 *The British Journal of Criminology* 1462; Ysabel Gerrard and Helen Thornham "Content Moderation: Social Media's Sexist Assemblages" (2020) 22 *New Media & Society* 1266 at 1278; Oliver Haimson and others "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas" (2021) 5 *Proceedings of the ACM on Human-Computer Interaction CSCW2* 466.

[6] Gartner "Gartner Forecasts Worldwide AI Spending to Grow 49.5% in 2026" (Press Release, 16 September 2026).

[7] "Meet Your New Assistant: Meta AI, Built With Llama 3" (18 April 2024) *Meta Newsroom* <https://about.fb.com>; "Announcing Grok" (3 November 2023) *xAI*, <https://x.ai/news/grok>.

adoption has coincided with a surge in deepfake incidents in 2025, marking a volatile shift in content authenticity.[8]

A particularly concerning development has been the rise of nonconsensual sexualised deepfakes.[9] For example, in 2026 users of X exploited Grok to generate nude or sexualised images of real women from publicly available photographs, prompting regulatory scrutiny in multiple jurisdictions.[10] Similar concerns arose from the widespread promotion of "nudifier" applications on Meta platforms, which generate fake nude images from uploaded photographs.[11] These developments illustrate how GenAI tools can facilitate image-based sexual abuse at unprecedented scale.

Concerns also extend beyond deepfakes. Reports have alleged that some GenAI chatbots on social media platforms generated sexualised interactions involving children,[12] while AI companions such as Grok's "Ani" have been criticised for embedding sexualised, racist and submissive female personas into everyday platform interactions.[13] While intended as entertainment, these avatars perpetuate the same patriarchal attitudes that underlie sexual harassment online, modelling interactions in which male users can exert control over digital representations of female characters.[14] Such features arguably normalise the commodification of femininity and extend patterns of gendered exploitation into virtual and everyday platform interactions.

AI systems are shaped by the social relations in which they are developed and used. Technology does not emerge independently of those relations, rather, decisions about what a system should do, what data it should be trained on, which users it is designed for and which risks warrant intervention are made within existing social and economic structures. Those choices can reproduce

---

[8] *Identity Fraud Report 2025–2026* (Sumsub, 20 January 2026).
[9] Russell Hawkins and Brent Mittelstadt "Deepfakes on Demand: The Rise of Accessible Non-Consensual Deepfake Image Generators" (2025) *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* 1.
[10] Emanuel Maiberg "Elon Musk's Grok AI Will 'Remove Her Clothes' In Public, On X" (6 May 2025) *404Media* < https://www.404media.co>; "eSafety raises concerns about misuse of Grok to generate sexualised content" (9 January 2026) *eSafety Commissioner* https://www.esafety.gov.au; "Ofcom launches investigation into X over Grok sexualised imagery" (12 January 2026) *Ofcom* <https://www.ofcom.org.uk>.
[11] "Spam and deepfake nudes: In which a swarm of X accounts floods the platform with spammy posts promoting an AI nudifier app" (13 June 2025) *Conspirador Norteño* <www.conspirator0.com>.
[12] Jeff Horwitz "Meta's AI Rules Have Let Bots Hold 'Sensual' Chats with Children" (14 August 2025) *Reuters* < https://www.reuters.com>.
[13] Jul Parke "Grok 4's new AI companion offers up 'pornographic productivity" (15 August 2025) The *Conversation* <https://theconversation.com/>.
[14] Parke, above n 13.

existing inequalities through the operation of the technology itself. Once deployed, the technology can also alter the conditions in which those social relations are experienced and reproduced. This reciprocal relationship is central to Judy Wajcman's concept of technofeminism.[15] This is important in the context of social media because the gendered effects of these systems cannot be separated from the conditions in which platforms are designed and operated. In other words, AI systems integrated in social media platforms are socio-technical systems that reflect and reinforce existing patriarchal power relations.[16] Where those relations underpin gender-based violence, technological systems can create new conditions in which that violence is reproduced. The emergence of nonconsensual sexualised deepfakes provides a clear example.

The connection between technology-facilitated abuse and existing forms of gender-based violence can be understood through Liz Kelly's concept of a continuum of sexual violence.[17] Kelly uses the continuum to show how forms of unwanted sexual conduct are often separated into different categories in law and social understandings. Some conduct is formally recognised as criminal, while other conduct is treated as an ordinary part of women's lives. McGlynn, Rackley and Houghton extend this analysis through a continuum of image-based sexual abuse, arguing that the legal and scholarly focus on "revenge porn" obscures the wider range of ways in which intimate images may be created or shared without consent.[18] Their analysis shifts attention away from the perpetrator's motive and towards the absence of consent and the control exercised over a woman's sexual representation.

Sexualised deepfakes sit within this continuum. The creator uses a woman's likeness to produce a sexualised representation of her body without her consent. The image may then be shared, reposted or reproduced by others, extending the loss of control. The harm can therefore begin with creation and continue through subsequent uses of the image.

The harm is also embodied. A deepfake appropriates the features through which a woman is recognised and places them into a sexual representation over which she has no control. Its synthetic

[15] Wajcman, above n 3.
[16] See for example, Forsch-Villaronga and others "Gendering Algorithms in Social Media" (2021) 23 *ACM SIGKDD Explorations Newsletter* 24; Karen Hao "Facebook's Ad-serving Algorithm Discriminates by Gender and Race" (2019) *MIT Technology Review* <https://www.technologyreview.com/>.
[17] Liz Kelly "The Continuum of Sexual Violence" in Jalna Hanmer and Mary Maynard (eds) *Women, Violence and Social Control* (Palgrave Macmillan, London, 1987) 46-60.
[18] Clare McGlynn, Erika Rackley and Ruth Houghton "Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse" (2017) 25 *Feminist Legal Studies* 25.

nature does not sever the connection between the image and the person depicted. Rather, that connection is central to the abuse. McGlynn and Toparlak describe the creation of sexualised deepfakes as the "new voyeurism", arguing that the creation itself can violate sexual autonomy even where the image is never circulated or the victim-survivor never becomes aware of it.[19] The wrong lies in the nonconsensual appropriation itself, rather than in what the creator subsequently does with the image or whether the person depicted ever discovers it. The possibility of sexualised deepfakes may also affect women who have never been targeted themselves. Knowing that an identifiable image can be transformed into a sexualised representation without consent creates an anticipatory threat that may influence how women present themselves and participate online.[20]

The technological conditions in which this conduct occurs are also gendered. GenAI systems are trained on existing cultural material, including representations shaped by longstanding patterns of gendered objectification. Research examining text-to-image models suggests that these patterns can be reproduced in model outputs. In one study of ten Stable Diffusion models, prompts concerning women produced 69.2% of the sexual or Not-Safe-For-Work images, compared with 30.8% for prompts concerning men. The models also generated female nudity more completely than male nudity, which the authors suggest may reflect the greater prevalence of female nudity in training data.[21]

The data from which these systems are developed can itself include images of women's bodies obtained without consent. GenAI can then draw on that material to produce new synthetic representations. Recent litigation against xAI illustrates the potential implications. A complaint filed by a child sexual abuse survivor alleges that Grok used images from her previous abuse to generate new synthetic depictions of her, renewing material that had already circulated for years.[22] The allegation illustrates a significant change in the conditions of abuse: material that was once finite can potentially be used as the basis for generating further representations of the same person.

---

[19] Clare McGlynn and Rüya Tuna Toparlak "The 'new voyeurism': criminalizing the creation of 'deepfake porn'" (2025) 52 *Journal of Law and Society* 204 at 211-212.

[20] While not about adult women, the threat of being "deepfaked", particularly in a sexual context, has led to some schools to make the decision not to post images of children to their public facing websites or social media pages, see: Karanama Ruru "School to stop posting children on their Facebook - should everyone follow suit?" (9 September 2026) *Stuff* <https://www.stuff.co.nz/>.

[21] Matthias Schneider and Thilo Hagendorff "Investigating toxicity and Bias in stable diffusion text-to-image models" (2025) 15 *Nature Scientific Reports*.

[22] Nick Robins-Early "Child sexual abuse survivor alleges Elon Musk's AI chatbot used photos of her to generate new illegal images" (3 September 2026) *The Guardian* <https://www.theguardian.com/>.

The resulting harm can therefore extend beyond the individual image to the continued technological reproduction of the person's sexualised identity.

Design choices further shape what users can produce. Content filters, safety measures and other guardrails determine whether sexualised representations of identifiable women can be generated and how easily that can occur. Where systems permit this conduct without effective safeguards, existing forms of gendered objectification are reproduced through the technological system. The user gains the ability to determine how a woman's likeness is sexually represented, while the woman herself has no control over that representation.

The prevalence of these uses indicates that such misuse cannot be treated as an isolated possibility. One study of deepfake videos found that 98% were pornographic and 99% of those depicted women.[23] McGlynn and Toparlak also identified substantial use of dedicated deepfake services, including one website attracting around 14 million users each month.[24] These patterns show how existing demand for the sexualisation of women intersects with technologies that make the production of sexualised representations increasingly accessible.

GenAI consequently changes the conditions in which image-based sexual abuse can occur. Sexualised representations can be produced without physical access to the woman depicted and potentially without her knowledge. The result is a form of abuse in which control over a woman's sexual representation can be exercised through technological systems shaped by the same gendered relations that sustain the underlying practice.

These developments raise questions about how New Zealand law can respond as AI-enabled gendered harm becomes normalised through its integration into widely used social media platforms. The following section uses the Grok Scenario as a case study to examine the capacity of the existing legal framework to respond.

1 How Well can New Zealand Respond? A Deepfake Case Study – the "Grok Scenario"

This analysis has attempted to situate the integration of AI into social media within its economic and gendered context to demonstrate that these technologies are products of structural incentives. Practices such as the proliferation of "nudifier" app ads on Meta platforms or the deployment of

[23] Security Hero "2023 State of Deepfakes: Realities, Threats, and Impact" (2023) <https://www.securityhero.io/>.
[24] McGlynn and Toparlak, above n 19, at 209-210.

AI chatbots capable of generating sexualised content illustrate that gendered harms are structurally (and unconsciously) woven into the design and deployment. This structural character complicates accountability. When harm arises, who is responsible? Is it the users maliciously prompting the AI tools? The failure of AI developers to provide sufficient guardrails? Or both? What of the wider concerns about the training of AI models or the normalisation of sexualised deepfakes as default on popular platforms? These dynamics raise critical questions about whether legal frameworks developed to address discrete acts of individual wrongdoing, as in New Zealand, can respond to harms that are technologically mediated and systemic in origin.

It is against this backdrop that the remainder of this article turns to a case study. To analyse how well the New Zealand legal framework can respond to emerging AI-facilitated gendered abuse via social media platforms, we will use an example of such abuse that has occurred recently: the use of Grok on X to "undress" another user without their consent. This example is particularly instructive because it involves the direct targeting of a real person, the nonconsensual generation of sexualised content (without being explicit), and the public dissemination of that content on a social media platform. It is also an example where an AI tool has been flagged multiple times for allowing users to generate such images, making this a recurring and foreseeable risk.

Using a real example of one user's experience,[25] and assuming all participants reside in New Zealand, the steps taken to achieve this can be summarised as the following:

1. A woman "person A" posts an image of their clothed body publicly on X
2. Person B, who is a stranger to Person A, "replies" to Person A's post using a prompt that engages Grok, X's AI chatbot
3. The prompt is "@Grok, undress her"
4. Grok generates and publicly posts a deepfake image of Person A now appearing to be in revealing lingerie (this appears in Person A's replies)

This example demonstrates the direct link between AI functionality and gendered harm, making it an ideal lens through which to interrogate the adequacy of existing law. While not an explicit image, the deepfake is an arguably sexualised image. It frames her body as an object for sexualised

[25] Maiberg, above n 10.

attention, reinforcing norms of female availability and desirability, and therefore produces similar harms to those associated with nude deepfakes. The fact that she appears in lingerie rather than nude does not remove the sexualised character of the image. Sexualisation can occur through the way a woman's body is presented and appropriated for sexual attention without her consent. The deepfake therefore interferes with her sexual autonomy by placing her embodied identity into a sexualised representation over which she has no control.[26]

By using this case study as a focal point, the discussion interrogates whether New Zealand's existing legal framework rooted in assumptions of individual liability and reactive remedies can meaningfully hold Person B or the platform accountable, or whether its limitations reveal a deeper misalignment between law and the socio-technical realities of AI integration on these platforms. This case study will be referenced throughout as the "Grok Scenario".

The following analysis will also be useful when considering nonconsensual sexualised content created by tools that are not integrated into social media platforms. With the proliferation of "nudifier" apps and websites across the internet,[27] and the general rise in popularity of AI chatbots and avatars that, while not purpose built for nudification, still have the capacity to produce such outputs, it is not the intention of this article to assert this is a problem with one single tool. Rather, it is to illustrate the readiness of New Zealand's legal framework in addressing emerging gendered abuse facilitated by AI.

III Applying New Zealand's legal and regulatory framework – Individual Perpetrator

In this section, we examine the application of New Zealand's legal and regulatory framework to the actions of Person B in our Grok Scenario. As it stands, there are limited options that could be used to pursue remedies against an individual perpetrator. One option is the Harmful Digital Communications Act 2015 ("HDCA"). Sections 22 and 22A, in theory, provide the Crown or Person A with the ability to pursue criminal charges against Person B, while s 19 provides for a civil regime. However, as we discuss below, both the criminal and civil regimes have significant limitations. It is evident that the HDCA was not designed with deepfake technology in mind,

[26] See McGlynn and Toparlak, above n 19.
[27] Margi Murphy "'Nudify' Apps That Use AI to 'Undress' Women in Photos Are Soaring in Popularity" (8 December 2023) *TIME* <https://time.com/>.

technology that is both ubiquitous and remarkably accessible. The proposed Deepfake Digital Harm Bill could provide a more explicit legal framework for prosecuting cases and updating the HDCA. This new legislation has the potential to be a fit-for-purpose solution, specifically addressing and incorporating harmful sexualised deepfakes within the scope of the HDCA's language. However, it is not without its own limitations.

Beyond the scope of the HDCA, a limited number of privacy torts and defamation may offer some recourse. However, much like the HDCA, these remedies appear to be ill-suited to the issues at hand.

These gaps in legal coverage necessitate the urgent development of new, specialised regulations designed to address the unique complexities of AI-enabled harm and ensure robust protection for individuals.

1 Harmful Digital Communications Act 2015

The HDCA is New Zealand's primary statutory response to serious online abuse, aiming to deter harmful digital communications and provide victims with effective redress.[28] It establishes communication principles guiding harmful content,[29] creates criminal offences for posting harmful communications and non-consensual intimate visual recordings,[30] and provides a civil regime enabling courts to order remedies where serious emotional distress is caused.[31]

(a) Section 22 (posting harmful digital communication)

Section 22 criminalises posting digital communications where three elements are met: (i) the poster intended to cause serious emotional distress to the victim, (ii) the communication would cause serious emotional distress to a reasonable person in the victim's position, and (iii) it did in fact cause serious emotional distress.[32] The statute also directs courts to consider a range of contextual factors when determining whether the post would cause serious emotional distress to a reasonable person, such as the anonymity of the poster, whether the material is false, the scale of dissemination,

[28] Harmful Digital Communications Act 2015 (HDCA), s 3.
[29] HDCA, s 6.
[30] HDCA ss 22, 22A.
[31] HDCA, s 19.
[32] HDCA, s 22(1). "Harm" is defined in s 4 as "serious emotional distress".

and the age of characteristics of the victim.[33] There are some notable normative hurdles when prosecuting Person B under s 22.

Section 22 requires proof that the defendant intended to cause serious emotional distress, that the communication would cause such distress to a reasonable person in the victim's position, and that it in fact did so. In the Grok Scenario, intent may be contested: a defendant could claim they did not know what the prompt would produce or characterised the conduct as a joke.[34] In New Zealand's first and only prosecution of someone for posting nonconsensual sexualised deepfakes, the defendant pleaded guilty to a s 22 offence after having confessed, clearly indicating his intent to cause the victim serious emotional distress.[35] However, not all victims will be so fortunate.

The reasonable-person inquiry also raises a gendered concern because context may include the platform on which the communication appears.[36] X's tolerance of sexual or controversial content could therefore be relied on to argue that the conduct was less unexpected or harmful.[37] Whether that argument should extend to sexualised manipulation of a user's image is murky. Many people use X for ordinary personal and professional purposes, and the platform's tolerance of sexual or controversial content does not necessarily establish an expectation that a user's own image may be manipulated in this way. These uses were not anticipated by the Law Commission in 2012. However, the implications of "context" bearing on the reasonable person inquiry could mean that the normalisation of abusive conduct on a platform risks becoming part of the basis for treating that conduct as less harmful. The more frequently users encounter sexualised content and other forms of gendered abuse on a platform, the easier it may become to characterise such conduct as an ordinary feature of participation in that space.

---

[33] HDCA, s 22(2).

[34] A claim of "just joking" is a common and somewhat successful barrier to holding individuals accountable for harmful online behaviour: Ariadna Matamoros-Fernández, Louisa Bartolo and Luke Troynar "Humour as an Online Safety Issue: Exploring Solutions to Help Platforms Better Address this Form of Expression" (2023) 12 *Internet Policy Review* 1.

[35] Gill Higgins "NZ makes first deepfake porn prosecution, but are we equipped for AI onslaught?" (17 October 2025) *1News* <https://www.1news.co.nz/>.

[36] The Law Commission explained that "context" is in reference to the digital platform in which the digital communication appears, as opposed to interpersonal context see Law Commission *Harmful Digital Communications: The adequacy of the current sanctions and remedies* (Ministerial Briefing Paper, 2012) 84.

[37] X Safety "Freedom of Speech, Not Reach: An update on our enforcement philosophy" (17 April 2023) *X Blog* <https://blog.x.com> .

More fundamentally, s 22 makes legal recognition of the wrong contingent on serious emotional distress. Evidence of depression, suicidal ideation, medical consultation, prescription of medication, or significant disruption to daily life can support a finding of actual serious emotional distress.[38] Conversely, indications of only annoyance, embarrassment, or low-level upset, will generally be insufficient.[39] This approach suggests that judges tend to require tangible manifestations of distress, which may disadvantage victims who respond with anger or resilience rather than overt emotional suffering. This makes the legal significance of the conduct dependent on its effects on the particular victim, even though the underlying wrong lies in the nonconsensual sexualisation and loss of control over sexual representation. A victim may experience that conduct as a serious violation without exhibiting the forms of distress most readily recognised by the law. Section 22 therefore provides an important route to redress, but its harm threshold limits its capacity to recognise the full nature of sexualised deepfake abuse.

(b) Section 22A (posting an intimate visual recording without consent)

Section 22A at first blush appears to be the most relevant for responding to AI-enabled nonconsensual sexualised deepfakes. The textual scope of the provision is broad: an "intimate visual recording" is defined in s 4 as a visual recording (including a still image) that is made in any medium without knowledge or consent of the subject of the photo and falls into one of two categories. The first is where it is recorded in a place where the subject is reasonably entitled to expect privacy and that person is shown (a) in state of nudity, including partially exposed breasts or in undergarments like lingerie or (b) engaged in sexual activity, or (c) shown toileting, showering, or dressing/undressing. The second is for activities that are colloquially referred to as "upskirting" or "downblousing", where images are taken from underneath or through someone's clothing.

The statutory language that an image may be "made in any medium using any device" creates at least a plausible textual foothold for including synthetic images, either in whole or in part. Where an AI tool such as Grok is used to produce a new image simulating a person as naked or partially naked and that image is then publicly posted as a reply to that person's account, a court sympathetic to a purposive reading of the statute could treat Grok as "any medium" and the social media prompt

[38] See for example, *R v L'Estrange-Corbet-Hooper* [2023] NZDC 21052, [41].
[39] See for example *Police v Brookman* [2020] NZDC 26801, *V v Google LLC* [2019] NZHC 488.

as “any device”, and therefore the synthetic output as an “intimate visual recording”. Such an interpretation would then shift the inquiry to the existence of consent.

However, there are two difficulties with applying s 22A to deepfakes. First, s 22A was introduced to address the problem of non-consensual posting of genuine intimate recordings (so-called “revenge porn”), after victims struggled to obtain justice under s 22 because of its requirement to prove both an intention to cause serious emotional distress and the actual occurrence of that distress.[40] Although Supplementary Order Papers[41] sought to clarify that synthetic images would also be covered, the Select Committee had already declined to add reference to “fake porn”, leaving the statutory language focused on intimate visual recordings.[42]

Second, the HDCA adopts the Crimes Act definition of “intimate visual recording”, which requires the subject to be in a place where they have a reasonable expectation of privacy and, under the relevant limb, to be “solely clad in undergarments”. While this may encompass a Grok image depicting a person in lingerie, it is less clear whether it extends to a deepfake depicting a person in a bikini or partial undress. More fundamentally, applying the reasonable expectation of privacy test to a synthetic image is problematic because the first stage of that inquiry turns on the physical place in which the recording occurred,[43] whereas no physical recording takes place when an image is deepfaked.

Ultimately, where nonconsensual deepfakes are prosecuted under the general offence in s 22 rather than the specific image-based offence in s 22A, a problematic divide emerges for victims. Those whose intimate images are real are afforded an assumption of harm and a straightforward inquiry into consent under s 22A. By contrast, those whose images are generated must satisfy the court that the communication both would and did cause serious emotional distress, and that the defendant intended that outcome. The law inadvertently privileges the authenticity of the image over the authenticity of the harm. In doing so, it risks creating a hierarchy of victimhood that fails to

[40] Anna Leask “Revenge porn: Man who uploaded sex tape escapes charges - victim 'gutted'” (12 July 2021) *Otago Daily Times* <https://www.odt.co.nz/>.
[41] Now referred to as Amendment Papers.
[42] Supplementary Order Paper 2021 (83) *Harmful Digital Communications (Unauthorised Posting of Intimate Visual Recording) Amendment Bill (305-2)*; Supplementary Order Paper 2021 (103) *Harmful Digital Communications (Unauthorised Posting of Intimate Visual Recording) Amendment Bill (305-2);* Justice Committee Report, *Harmful Digital Communications (Unauthorised Posting of Intimate Visual Recording) Amendment Bill (305-2)* 4-5.
[43] *Keighran v R* [2022] NZCA 75.

recognise that the injury of nonconsensual sexualisation does not depend on whether the image was "real" but on the fact that it was unwanted.

This distinction resembles a broader pattern identified by feminist scholarship in relation to sexual violence, where some forms of sexual violation have historically been treated as more serious than others. The stereotypical stranger rape, for example, has often been understood as the clearest form of sexual violence, while rape by an acquaintance or intimate partner has historically been more readily minimised.[44] The circumstances surrounding the violation can consequently influence perceptions of its seriousness even though the central issue remains the absence of consent. A similar hierarchy can arise here when the authenticity of the image becomes a proxy for the seriousness of the violation.

(c) Deepfake Digital Harm and Exploitation Bill

A members bill proposing to expressly include deepfakes in the definition of "intimate visual recording" in both the Crimes Act and the HDCA passed its first reading in May 2026.[45] This Bill would expressly criminalise the making, possessing, publishing, importing, exporting, selling or posting online of sexualised deepfakes without knowledge or consent of the subject.

The bill proposes to insert into the definition of "intimate visual recording" the following:[46]

> intimate visual recording includes a visual recording that has been created, synthesised, or altered without the knowledge or consent[47] of the person who is the subject of the recording, and appears to show the person
>
> (a) naked or with their genitals, pubic area, buttocks, or female breasts exposed, partially exposed, or clad solely in undergarments; or
>
> (b) engaged in an intimate sexual activity; or

[44] Susan Estrich *Real Rape* (Harvard University Press, Cambridge, 1987).
[45] Deepfake Digital Harm and Exploitation Bill 2025 (213-1).
[46] Deepfake Digital Harm and Exploitation Bill, cl 4(1)(1A) and cl 6(1)(ab).
[47] There are additional points to raise regarding the Bill that are not the focus of this article but see: Cassandra Mudgway "Deepfake Digital Harm and Exploitation Bill: Chasing the Technology" (2026) 1 *New Zealand Law Journal* 6.

(c) engaged in showering, toileting, or other personal bodily activity that 10 involves dressing or undressing.

This definition would remove the need to analyse whether there was a reasonable expectation of privacy, leaving a much more straightforward pathway for prosecution of Person B under s 22A in our Grok Scenario. It would also go some way in resolving the current gap for victims between s 22 and s 22A of the HDCA. Consent to one form of sexual representation does not amount to consent to another. A person may have voluntarily shared sexual or intimate content elsewhere and still be wronged when their likeness is appropriated to create a sexualised deepfake in an unrelated context. The relevant question is whether they consented to the particular use of their identity, rather than whether they have previously chosen to make sexualised images of themselves available. The proposed definition would also avoid potential hurdles posed by the requirement under s 22 of proof of specific intent and serious emotional distress.

Instructing Grok to "undress" Person A results in an output that either creates or synthesises an image. An image of Person A in lingerie would be “clad solely in undergarments” (a). In our scenario, Person A did not know or consent to its creation nor did they consent to having it posted. As Grok creates and then posts the image automatically, this seemingly fulfills the elements of s 22A. Moreover, by using Grok in this way, Person B potentially commits two offences at the same time. First, the making of the sexualised deepfake (s 216H of the Crimes Act) and then posting it on X without consent (s 22A HDCA).

Despite its clear advantages, the proposed definition is not without difficulty. First, the formulation that a recording “appears to show the person” introduces an evidentiary and conceptual ambiguity that may be open to contest. While the intention is to capture synthetic and altered images, the statutory language invites questions about recognisability, realism, and the perspective from which “appearance” is assessed. In practice, defendants may argue that an image is too stylised, exaggerated, or obviously artificial to be said to “appear” to show the victim. While images may look less "artificial" as GenAI models continue to improve, this wording still may not capture purposefully stylised deepfakes, such as the surge in models designed to create “waifu” images of anime-style women with exaggerated features.[48] In its current form, the provision risks allowing

[48] Laura Wagner and Eva Cetinic “Perpetuating Misogyny with Generative AI: How Model Personalization Normalizes Gendered Harm” (2025) *arXiv preprint*.

disputes about visual authenticity to obscure the central wrong: the nonconsensual sexualisation of a real person.

Secondly, the proposed definition continues to conceptualise harm through a narrow and traditional understanding of sexual intrusion. By linking liability to nudity or specific intimate behaviours, the law privileges certain "core" sites of sexual harm while marginalising other common forms of gendered sexualisation. This is relevant in the context of AI-generated content, where malicious outputs often involve swimwear or the selective removal or alteration of clothing that stops short of rendering the subject "clad solely in undergarments". Many of the Grok-generated examples that have circulated publicly fall into this grey zone, namely, highly sexualised, humiliating, and unwanted, however, potentially outside the statutory definition. The effect is to reproduce a familiar hierarchy in which only certain forms of sexual harm are legally recognised as harmful, despite feminist scholarship and victim testimony consistently demonstrating that harm does not map neatly onto degrees of nudity.[49]

Finally, the criminalisation of malicious individual behaviour, while necessary, captures only part of the wider deepfake ecosystem. The Bill does not address threats to create or share deepfakes, the use of harmful images in AI training, or the technological conditions that allow sexualised deepfakes to be produced so readily. Criminalisation may provide a response once a particular act has occurred, while the conditions that facilitate that conduct remain in place. While criminalising the creation and distribution of deepfakes may seem to deliver justice, it largely benefits victim-survivors who seek redress through the criminal process. Carceral responses seldom confront the structural and psychological dimensions of image-based sexual abuse and can retraumatise victims while excluding those who distrust or fear engagement with the criminal justice system.[50] Arguably, meaningful redress must also extend beyond punishment to include holistic, non-criminal pathways that centre healing, restoration, and prevention.

[49] Joanne Conaghan "Law, harm and redress: a feminist perspective" (2002) 22 *Legal Studies* 319; Kcasey McLoughlin and Alex O'Brien "Feminist Interventions in Law Reform: Criminalising Image-Based Sexual Abuse in New South Wales" (2019) 8 *Laws* 35.
[50] Glitch *Beyond the Content Takedown: Non-Criminal Redress for Intimate Image Abuse* (2025) 12-14.

(d) Section 19 - the civil regime

The HDCA's civil regime offers an alternative route to redress, although its practical accessibility and effectiveness in responding to harmful content warrant consideration. Person A may pursue civil remedies under s 19 of the HDCA. If successful, the court may make any number of different orders, such as orders for the poster or the platform to remove the harmful content, for the poster to cease communication to and about the applicant, or an order to reveal the identity of an anonymous poster. A complainant must first lodge a complaint with Netsafe and show breach of a communication principle causing serious emotional distress. Making an application under the HDCA, unlike many other civil routes, does not incur cost. However, there is no urgency attached to the application, which means there could be lengthy delays in having an application heard.

In the Grok Scenario, Principle 1 (a digital communication should not disclose sensitive personal facts) is unlikely to apply because a synthetic image does not disclose a "fact". Principle 3 may apply, however, if the creation and publication of a nonconsensual sexualised deepfake is considered "grossly offensive" to a reasonable person. As with the reasonable-person test under s 22, the contextual nature of this assessment creates a potential difficulty where the norms of the online forum are influential.[51] The gendered implications of this approach, discussed above, are equally relevant here, namely, where sexualised abuse has become normalised online, the prevailing online environment may itself influence the baseline against which offensiveness is assessed. Principle 4 (a digital communication should not be indecent or obscene) is more likely to apply where the image depicts nudity, sexual acts or a person solely clad in undergarments, while Principle 10 (a digital communication should not denigrate a person by reason of that person's … gender) may apply where the deepfake denigrates the complainant on the basis of gender, including through sexual humiliation or gendered stereotypes.

The District Court has a broad range of remedies, including take-down, cease-and-refrain, non-encouragement, identification, correction, apology and right-of-reply orders.[52] For a single post by an identifiable defendant, take-down and cease-and-refrain orders would likely be most useful, but these remedies become less effective once the image has been copied or redistributed beyond the

[51] *Ministerial Briefing Paper*, above n 36, at 117.
[52] HDCA, s 19.

defendant's control. The HDCA also provides no mechanism for damages, leaving a significant remedial gap where the image continues to circulate despite court orders.

2 The Privacy Torts

There are a few avenues for redress outside of the HDCA. Given that nonconsensual sexualised deepfakes intervene with an individual's likeness in a way that simulates them in a state of undress or in a sexual manner that many would consider private, it is instructive to consider the applicability of New Zealand's privacy torts. New Zealand courts have only recognised two causes of action for privacy breaches: "publicity of private facts" and "intrusion upon seclusion". These torts are designed to protect individuals from unauthorised exploitation of their personal information and intrusion upon their dignity. While the rise of GenAI may challenge traditional understandings of privacy,[53] the principles underlying these torts provide a useful framework for evaluating whether nonconsensual sexualised deepfakes constitute actionable invasions of privacy. The following analysis applies the established elements of each tort to our Grok Scenario, where Person A is the plaintiff and Person B is the defendant.

(a) Publicity of private facts

The tort of invasion of privacy by publicity of private facts, as articulated in *Hosking v Runting*,[54] requires two essential elements: (i) the existence of facts in respect of which there is a reasonable expectation of privacy; and (ii) publicity given to those private facts that would be considered highly offensive to an objective reasonable person. There is a defence of legitimate public concern. If successful, the most likely remedy will be damages.[55]

The tort concerns private facts, not public facts. It is generally agreed that there is less expectation of privacy regarding something that occurs in a public place.[56] On this view, the defendant could argue that the Grok deepfake should be characterised as a public fact because it was generated and disseminated within a public forum, where the plaintiff had already chosen to share her image. By posting on a globally accessible platform, the plaintiff arguably placed her likeness in the public

[53] Karina Devi Etminani "Generative Privacy Doctrine: The Case for a New Legal Privacy Framework For Gen^AI" (2025) 52 *UC Law Constitutional Quarterly* 305.
[54] *Hosking v Runting* [2005] 1 NZLR 1.
[55] *Hosking v Runting* at [149].
[56] *Hosking v Runting* at [119].

domain, rendering derivative uses of it *public* rather than *private*. Increasingly hosting more extreme content as well as sexual content, the setting of X may cast a shadow over what facts or information would have a reasonable expectation of privacy when one voluntarily engages on the platform even where one's naked body is shared. The defendant may go further and argue there is no privacy when using a public account on such a platform. This would be an untenable position as such an argument would open the door, for example, to allowing widespread doxxing[57] of individuals merely because that individual has an account on X. Moreover, where images are taken from elsewhere on the internet then posted to X without consent where they are then deepfaked, it cannot then be said a reasonable person should expect that to happen merely because X is a more extreme social media platform. What is considered "reasonable" has been critiqued by feminist legal scholars as obscuring women's lived experiences because the so-called neutrality of an objective reasonable person often defaults to a masculine view of the world.[58] It cannot be considered reasonable for women to expect to be subject to image-based sexual abuse. There is no "bright line", but there must be a line.[59]

Even within the most public spaces, courts have recognised privacy interests even in public settings where intimate information is involved, for example, an intimate conversation between husband and wife during the aftermath of a road accident.[60] Further, courts have repeatedly recognised that facts concerning the intimate parts of one's body fall squarely within the scope of privacy interests.[61] Therefore, the plaintiff would be in a position to argue that although there is less expectation of privacy concerning her likeness on a public forum like X, there is still an expectation of privacy as to parts of her body she decided not to show and to her intimate self. Notwithstanding shifting norms regarding the use of GenAI in recent years,[62] a reasonable person applying contemporary moral standards would likely still consider having a reasonable expectation of privacy regarding intimate parts of their body.

---

[57] Doxxing is the posting or sharing of private identifying information without consent.
[58] Ruth Gavison "Privacy and the Limits of Law" (1980) 89 *Yale Law Journal* 421; Catharine MacKinnon *Are Women Human?* (Harvard University Press, Cambridge, 2006).
[59] Per Gleeson CJ *Australian Broadcasting Corp v Lenah Game Meats Pty Ltd* (2001) 208 CLR 199, 226.
[60] *Andrews v Television New Zealand Ltd* [2009] 1 NZLR 220.
[61] *P v D* [2000] 2 NZLR 591; *Hosking v Runting*, above n 54, at [117].
[62] Andrea Pocol "As generative AI becomes more sophisticated, it's harder to distinguish the real from the deepfake" (26 March 2025) *The Conversation* <https://theconversation.com/>.

The distinction of what is public and what is private has long been criticised by feminist legal scholars as excluding gendered harm from legal recognition.[63] Digital spaces can also collapse the distinction, as it becomes normalised to share once typically private information on public platforms. Not only does the inquiry into what is reasonably expected to be "private" become more complicated, it is arguably unsuited to assess privacy breaches experienced by women and girls in digital environments. This is because "reasonable expectations" are assessed against norms of visibility and consent that inadequately capture gendered power relations and the sexually abusive dynamics shaping women's and girls' digital visibility. It is important to note that we are exploring a scenario where Person A is a private person, and these principles may not apply the same way if she were a public person[64] or a person who has shared intimate images online on private platforms, such as OnlyFans, but not on public platforms like X.[65] Generally, privacy torts concern the publicity of true facts as opposed to untrue facts. In our scenario, the deepfake is not a real depiction of the plaintiff. In *Hosking*, when discussing the difference between the tort of defamation and privacy, Gault P and Blanchard J surmised that the privacy tort was about emotional harm rather than reputational harm and that "[t]he objectionable disclosure may be entirely factually accurate."[66] Following from this dicta, it could be equally argued that publicity of false or untrue private information causing distress and humiliation would fall within the scope of this privacy tort.[67] Similarly, in *Hill v Radio One*, the Broadcasting Standards Authority held that "it is the quality of the information about a person which determines whether or not it is private, and not the veracity of the information."[68] This approach reflects the normative purpose of the privacy tort to safeguard dignity and autonomy. A fabricated intimate image can cause humiliation and loss of control over self-presentation, harms that the privacy tort is designed to address.

[63] Frances Olsen "The Myth of State Intervention in the Family" (1985) 18 *University of Michigan Journal of Law Reform* 835; Nicola Lacey "Theory into Practice? Pornography and the Public/Private Dichotomy" (1993) 20 *Journal of Law and Society* 93; Danielle Keats Citron "Sexual Privacy' (2019) 128 *Yale Law Journal* 1870.
[64] "Tiffany Salmond shares new evidence of explicit deepfakes amid online abuse" (25 October 2025) *NZ Herald* <https://www.nzherald.co.nz>; See also discussion in *Hosking v Runting*, above n 54, at [141].
[65] Maggie Dupré "OnlyFans Model Shocked After Finding Her Pictures With AI-Swapped Faces on Reddit" (16 May 2025) *Futurism* <https://futurism.com/>; Jason Koebler and Emanuel Maiberg "Inside the Booming 'AI Pimping' Industry: AI-generated influencers based on stolen images of real-life adult content creators are flooding social media" (20 November 2024) *WIRED* <https://www.wired.com/>.
[66] *Hosking v Runting*, above n 54, at [137].
[67] And where the damage is reputational, provide the basis for a claim in defamation (see below).
[68] *Hill v Radio One* BSA 2013-074, 4 March 2014 at [12].

The deepfake must be highly offensive to a reasonable person in the position of the plaintiff. There is a strong argument available to a plaintiff that based on the intimate and nonconsensual nature of the deepfake that it would be considered highly offensive to a reasonable person. There is likely no public interest in publicising a nonconsensual sexualised deepfake, so we will not discuss the defence.

(b) Intrusion upon seclusion

For a successful claim of intrusion upon seclusion, the plaintiff would need to prove (i) an intentional and unauthorised intrusion (ii) into seclusion, (iii) with which there is a reasonable expectation of privacy, and (iv) the intrusion was highly offensive to a reasonable person in the position of the plaintiff.[69]

The nonconsensual nature of the deepfake makes *unauthorised* a straightforward inquiry, but whether or not it is an *intentional* intrusion is less certain. As Grok was a newly deployed AI and implemented in a general social media platform, it could be arguable that Person B could not know whether a sexualised deepfake would follow. According to Whata J in *C v Holland*, the intrusion must be the result of an affirmative act as opposed to an accident.[70] The arguable action is the input of the prompt "@Grok, undress her" by the defendant. The alleged intrusion is the generated deepfake. The latter cannot occur without the former, so the "intentional" part of this element may be made out. It is perhaps analogous to opening a door to a changing room. A person cannot be certain that someone is changing behind the door when they open it, but it is an intentional act nonetheless when they happen to find someone there. Whether or not Person B knew for certain the "undress her" prompt would result in a sexualised deepfake, it is still an intentional act, whether one is created or not.

Is the Grok deepfake an intrusion into the plaintiff's seclusion? The notion of seclusion extends beyond physical spaces to encompass intimate personal activity, space or affairs, and can therefore capture digital intrusions. Like all privacy torts, it is concerned with the infringement of personal autonomy. Again, one's near-naked body is almost always considered intimate and thus shielded from the observation of others. But where does the intrusion occur? Is the generation of an image that looks like the plaintiff, but is not actually the plaintiff, an intrusion? Much like publicity of

[69] *C v Holland* [2012] 3 NZLR 672 at [94].
[70] *C v Holland* at [95].

private facts, to answer this the court will need to consider incremental development of the tort to reflect purely digital contexts.

Traditional intrusion cases presuppose an act of observation or capture: the wrong arises because the defendant has seen or recorded something private that they were not authorised to access. Deepfakes, by contrast, are generated rather than observed, no private material is "obtained" in any conventional sense. Arguably, the harm is functionally equivalent to being secretly observed or filmed naked, because in both instances the individual loses control over the boundaries of their intimate identity. A deepfake is created to be looked upon, and thus there is an intrusion.

As the final elements (expectation of privacy and intrusion being highly offensive to a reasonable person) are the same as publicity of private facts, similar arguments may be raised here. Courts in *C v Holland* and *TJ v Television New Zealand Ltd* confirmed that this assessment must consider the degree, context, and motive of the intrusion;[71] here, relevant factors pointing towards the intrusion being highly offensive include the deliberate nonconsensual sexualisation and public posting of the image.

(c) Remedies

The practical hurdle for a plaintiff in our Grok Scenario will be the limited available remedies relative to the potentially significant financial cost of bringing a civil case, the financial position of the defendant to afford damages and costs if awarded, and the time in litigating the dispute at trial. For privacy torts in New Zealand there are likely two remedies: injunctions or general damages. An injunction to prevent publication would not be practicable because the deepfake has already been posted, leaving the plaintiff to claim damages.

An early privacy case involving the publication of photographs of a plaintiff's genitals without consent resulted in an assessment of damages that took into consideration both the plaintiff's and defendant's behaviour and motivations, the plaintiff's humiliation and the publicity's impact on her dignity, and how identifiable she was in the photographs.[72] She was awarded $2,500. In *Peters v Bennett*,[73] Venning J reviewed the case law in NZ and the UK to determine the appropriate levels of damages, suggesting that an award between $75,000 and $100,000 was justified based on a

[71] *C v Holland* at [96] and *TJ v Television New Zealand Ltd* BSA 2013-092, 17 June 2014 at [22]-[24].
[72] *L v G* [2002] DCR 234, 249-250.
[73] *Peters v Bennett* [2020] NZHC 761.

deliberate breach of privacy with an intention to cause embarrassment and distress.[74] Ultimately, damages for privacy torts is about the measurement of intrusion upon a person's dignity.[75]

The factors that would impact on the provision of damages and the final amount of the award could be the following: the defendant could argue that the plaintiff's decision to post on X, a platform widely associated with minimal moderation and explicit content, should mitigate damages. This argument positions the plaintiff as having assumed a foreseeable risk of privacy harms, thereby reducing the seriousness of the invasion. Conversely, voluntary engagement on a public platform does not amount to consent for nonconsensual sexualisation. Courts have consistently rejected the notion that contextual publicity strips individuals of *all* privacy rights, and the same reasoning applies to damages.[76] However, the defendant may still persuade the court that this factor warrants a modest downward adjustment in quantum, particularly when compared to cases involving private or confidential settings. This would be a problematic argument. Treating platform participation as a factor reducing damages risks treating women's decision to participate in public life as an assumption of risk for subsequent sexual abuse. That approach would be particularly difficult to justify where the very harm complained of consists in taking control of how a woman's likeness (of her body) is used. The concern extends beyond the individual claimant. If women are expected to limit their use of platforms to reduce the risk of being deepfaked, the practical cost of privacy protection is imposed through reduced participation in online spaces.

The plaintiff could emphasise that the deepfake's sexualised nature is an aggravating factor that should elevate damages. Courts have historically treated intimate material as engaging the highest order of privacy interests, even though the image is not explicit, it nonetheless impacts on the personal autonomy of the plaintiff. This factor supports an award at the upper end of the scale, aligning with judicial observations that damages should reflect the gravity of the affront to personal dignity rather than the mere fact of dissemination.

There may be some persuasive arguments available to the defendant regarding whether the deepfake was the result of deliberate action and an intention to humiliate. While the defendant

[74] *Peters v Bennett* at [275].
[75] *Henderson v Walker* [2019] NZHC 3020 at [17]–[18], see also *Hosking v Runting*, above n 54, at [239], [258], per Tipping J. Notwithstanding other surrounding harm that might also be considered, such as financial loss resulting from the invasion of privacy, which falls outside the scope of this article.
[76] See *Hyndman v Walker* [2021] 2 NZLR 685; *Driver v Radio New Zealand Ltd* [2020] 3 NZLR 76; *Rickard v Television New Zealand Ltd* BSA 2016-098, 19 April 2017.

initiated the process by prompting the AI tool with the instruction "undress her" the actual creation and automatic posting of the image on X was facilitated by Grok's design and the absence of effective guardrails in the AI system. This raises questions about foreseeability and intent. If the defendant knew that the tool would generate and publish a sexualised image, or if the prompt was motivated by a desire to humiliate, this would weigh heavily in favour of higher damages. Conversely, if the defendant believed the AI would refuse the request or respond with a humorous or non-sexualised alternative, the conduct may be characterised as reckless rather than malicious, potentially moderating the award. This interplay between human agency and technological automation underscores the complexity of attributing culpability in AI-mediated privacy breaches and may influence the normative assessment of damages.

Overall, from the plaintiff's perspective, damages are a blunt and retrospective remedy for a harm that is both immediate and ongoing. Monetary compensation cannot restore control over one's image once it has entered networked circulation, nor does it address the chilling effects such violations can have on women's participation in online spaces. The gap between the lived experience of harm and the remedial response offered by privacy torts highlights a structural mismatch between the nature of image-based sexual abuse and the remedial logic of private law.

3 Defamation

Another relevant cause of action in New Zealand is defamation. Unlike the privacy torts, defamation is not about humiliation and distress but harm to reputation. To establish a claim in defamation, three elements must be proved: (i) that the statement was defamatory; (ii) that the statement is about the plaintiff; and (iii) that the statement was published by the defendant. A statement is defamatory if it tends to lower the plaintiff in the estimation of right-thinking members of society, if it tends to make people shun or avoid the plaintiff, or is a statement that brings her into disrepute.[77]

A deepfake may constitute a "defamatory" statement where it conveys a false and damaging representation of a person. Defamation may arise from any form of communication, whether written, spoken, or visual. Because the falsity is inherent in the synthetic nature of the image, and recognition of the subject is often immediate, a realistic and widely shared deepfake could well

[77] *Karam v Parker* [2014] NZHC 737, [25].

satisfy all three elements of defamation: falsity and harm to reputation, reference to the plaintiff, and publication to third parties. However, deepfakes that merely simulate the plaintiff being naked or in a state of undress may not necessarily be considered "defamatory". It is not merely that the image is a false representation, but it must also reflect adversely upon the plaintiff so that their standing among right-thinking people is lowered.[78] If the deepfake was posted alongside claims that the plaintiff was an adulterer, or if the deepfake depicted the plaintiff engaging in criminal or immoral conduct (in the eyes of right-thinking people) then this may be "defamatory". In our Grok Scenario, however, it is unlikely that a deepfake simply simulating Person A being in lingerie is in and of itself "defamatory".

In sum, Part III has illustrated several legal hurdles and gaps regarding pursuit of Person B for criminal or civil remedies. Part IV turns our attention to the platform in which the deepfake was both created and posted.

IV Applying New Zealand's legal and regulatory framework – The Platform

This section will examine how New Zealand's legal and regulatory framework may apply to a platform that deploys GenAI (such as X) that is subsequently used to produce harmful sexualised deepfakes. There is a crucial distinction to be made between X Corp and xAI, which highlights a peculiarity that applies to online platform regulation. X Corp is merely a deployer of Grok, whereas xAI is the developer; in practice, both organisations are owned by the same individual. This separation provides a layer of obscuration which significantly complicates the victim's ability to seek recourse. This is because X Corp, as a deployer, is not as proximate to the harm as the developer or individual who prompted the GenAI tool. Such separations are common in AI development and deployment.

The separation between developer, deployer and user also has consequences for how gendered harm is legally understood. Each actor occupies a different position within the technological system, while the harmful outcome depends on their interaction. Person B supplies the prompt, xAI develops the model, and X incorporates the system into a platform through which the resulting image is generated and disseminated. The legal framework tends to examine these contributions separately. That division can obscure where the capacity for prevention sits. The woman subjected

[78] *Sim v Stretch* [1936] 2 ALL ER 1237 (HL).

to the deepfake has no control over the model, its safeguards or the decision to integrate it into X. X and xAI, by contrast, have significant control over those features before a particular user initiates an abusive act. The distribution of legal responsibility therefore does not necessarily correspond with the distribution of technological power.

If Person A sought to target the platform where the sexualised deepfake was posted, they will likely have four avenues to pursue. The first is through New Zealand's voluntary *Code of Practice for Online Safety and Harms*, the second through the Privacy Act 2020, third under defamation law, and the fourth through a finding of negligence by that platform. However, as discussed below, it is difficult to map offending of this kind cleanly to any of these options, making a satisfactory outcome for Person A unlikely in practice.

1 The Code of Practice for Online Safety and Harms

The *Code of Practice for Online Safety and Harms* ("the Code") was drafted in 2022 by Netsafe and NZTech with major platforms, including X Corp.[79] The Code provides a voluntary co-regulatory framework through which companies commit to high-level principles on harm reduction, moderation, and transparency. In terms of compliance, the Code is enforced by an Oversight Committee. Where companies fail to comply with the Code, the Committee may issue a warning to the company alongside recommended steps for compliance, at worst the Committee may terminate the company as a signatory to the Code.

Despite its broad scope, the Code makes no reference to image-based sexual abuse, deepfakes, or any form of AI-generated harm. This is not surprising, because GenAI on social media platforms have only recently been implemented. It is surprising, however, that gendered abuse is not specifically named. The Code does acknowledge sex and gender identity but only in its definition of hate speech, rather than as a basis for harassment. The absence of image-based sexual abuse is also notable given the prevalence of so-called "revenge porn" by the 2020s.[80] This reflects a longstanding pattern whereby abuse disproportionately experienced by women are frequently treated as subsidiary examples of more general forms of wrongdoing rather than recognised as distinct harms requiring specific attention. Just as domestic violence and sexual harassment were

[79] *Aotearoa New Zealand Code of Practice for Online Safety and Harms* (July 2022).
[80] See Nicola Henry, Nicola Gavey, Clare McGlynn and Erika Rackley "'Devastating, like it broke me': Responding to image-based sexual abuse in Aotearoa New Zealand" (2023) 23 *Crim & Crim Justice* 861.

historically obscured by ostensibly gender-neutral legal frameworks, image-based sexual abuse risks becoming invisible when subsumed within broad and undifferentiated categories of online harm. The absence of any express recognition of gender-based abuse therefore shapes what forms of harm become visible within the Code's governance framework and, by extension, what risks platforms are encouraged to prioritise.

The Code's limitations are structural. Its commitments are aspirational, framed in language such as "seek to" or "support," which signals effort rather than obligation.[81] This model was always a form of soft accountability, but it is inadequate where harm arises from the design of AI tools, not just user behaviour. The Code contains no minimum safety standards, no duty to test AI tools for foreseeable misuse, and no requirement to conduct risk assessments, let alone gendered or algorithmic ones. Under the current framework, a platform can release an AI system capable of producing nonconsensual sexualised deepfakes on command and still claim full compliance.

As such, holding X Corp accountable to the Code in our Grok Scenario is not a viable option.

2 The Office of the Privacy Commissioner

Person A may be able to make a complaint to the Office of the Privacy Commissioner ("OPC") alleging X Corp interfered with their privacy contrary to the Privacy Act 2020. Section 70 allows the OPC to receive complaints from individuals who allege an agency has interfered with their privacy. Although X Corp is an overseas company, its services still operate within New Zealand and are therefore subject to the Privacy Act.[82] The privacy information in question is person A's likeness.[83]

The OPC has not yet issued specific guidance on the application of the Privacy Act to sexualised deepfakes. However, there is broader guidance on *AI and the Information Privacy Principles* that makes clear the definition of "personal information" is intentionally broad and technologically neutral.[84] In particular, the OPC explicitly states that personal information includes inaccurate or made-up information, including deepfake images.[85]

---

[81] See Cassandra Mudgway and Sarah Bickerton "Hashtag Online Misogyny: A Proposal For Online Safety Regulation In Aotearoa New Zealand" (2026) 10 *New Zealand Women's Law Journal* 189.
[82] Privacy Act 2020, ss 8-9.
[83] Privacy Act 2020, s 7 meaning of "personal information".
[84] Office of the Privacy Commissioner, *AI and the Information Privacy Principles* (2023).
[85] Office of the Privacy Commissioner, *AI and the Information Privacy Principles*, at 1.

To amount to inference with privacy under the Privacy Act, Person A must show (i) a breach of one or more of the Information Privacy Principles (IPP) and (ii) that the breach caused harm i.e. serious humiliation, loss of dignity, or harm to reputation.[86] As discussed above, although unlikely to be defamatory harm, the deepfake would likely be found to cause serious emotional distress, thus, be considered "harm" under the Privacy Act. The critical legal question becomes whether Grok or X has collected,[87] used,[88] or disclosed[89] Person A's personal information through the creation or hosting of the deepfake.

Where a model's training data includes images scraped from the public internet, including images of the complainant, the creation of the deepfake arguably involves the *use* of that information for a purpose unrelated to its original collection (IPP 10). Even if training data is not directly invoked at inference time, the model's capacity to reproduce recognisable features may still be characterised as a form of use under the Privacy Act. However, because the deepfake is in fact a new image, the resulting near-naked depiction is not a representation of the data that it holds. Nevertheless, according to the OPC guidance on *AI and the Information Privacy Principles*, using information as a prompt (like Person A's image) or generating new information about a person (even if it is false or inaccurate) is "collecting" personal information.[90] If deemed to have collected personal information by Grok being used this way, then a nonconsensual sexualised deepfake would most certainly be unfair and therefore breach IPP 4 (collection of information must be legal and fair).

When a deepfake is disseminated on X, the platform's systems "disclose" the content by making it publicly accessible (IPP 11). X Corp is likely to argue that any "disclosure" is performed by the user who uploads the content, not the platform.[91] This would implicate either Person A or Person B. On this view, the platform merely hosts third-party content and does not "use" or "disclose" personal information in the manner contemplated by the Privacy Act. Given the aforementioned purposefully broad definition of "personal information", this is not a strong argument.

---

[86] Privacy Act 2020, s 22.
[87] Privacy Act, s 22 IPP 1, 2, 4.
[88] Privacy Act, s 22 IPP 10.
[89] Privacy Act, s 22 IPP 11.
[90] Office of the Privacy Commissioner, *AI and the Information Privacy Principles*, above n 84, at 5.
[91] Elon Musk argued that individual users who prompt Grok to create deepfakes will be booted from the platform, but did not take responsibility for failing to include appropriate guardrails to prevent them: "XSafety statement" (4 January 2026) <https://x.com/>.

The most likely hurdle for Person A in showing a breach of the Privacy Act (or in the OPC's investigation of the alleged breach) is causation. Who is collecting, using and disclosing information in our scenario? There is no doubt that Person B does so (according to the OPC guidance), but what about X Corp? This would depend on what Grok or X Corp does with the information. If the information (prompt or resulting deepfake) is retained by Grok, then xAI (as developer) would be deemed to be "collecting" information and IPP 1, 2, 4, 11 are all engaged. As the information (prompt and deepfake) are processed on X, then, according to the OPC guidance, X Corp would also be considered as "collecting" information. This is likely, as in order to host content on X, then the deepfake would need to be stored on X for it to appear in the X thread. This illustrates a tension within the Privacy Act's technologically neutral approach, namely, the law can recognise a synthetic image as personal information, while responsibility for its creation and processing remains divided between the different actors involved in the technology.

Once a complaint is accepted, the OPC may open a formal investigation under Part 5 of the Privacy Act. This may include the OPC requesting information from X Corp or xAI about how Grok operates. Given the company's aversion to being transparent with overseas regulators, such as the Australia's eSafety Commissioner,[92] this might be difficult to achieve. The OPC can issue compliance notices to X if they fail to cooperate, which may then be enforced by the Human Rights Review Tribunal ("HRRT").[93] Ignoring the HRRT is a criminal offence which can result in a $10, 000 fine.[94] However, given X Corp is currently valued into the billions of dollars,[95] the fine is laughable.

Whether or not X Corp cooperates with an OPC investigation, or where a determination of a breach of the IPPs is made, the Commissioner may refer the matter to the HRRT for determination and, if Person A is successful, issue damages in her favour.[96] Damages are compensatory only, and can be awarded for "humiliation, loss of dignity, and injury to the feelings of the aggrieved

[92] Rebecca Picciotto "Tech Australia fines X, formerly Twitter, for not answering questions on child abuse content" (15 October 2023) *CNBC* <https://www.cnbc.com>; eSafety Commissioner "Full Federal Court rejects X Corp. appeal" (media release, 31 July 2025).
[93] Privacy Act 2020, ss 123-129.
[94] Privacy Act 2020, s 133(3).
[95] Justin Baragona "X finally valued at $44 billion, the exact sum Elon Musk paid for the social media site in 2022" (*Independent* 19 March 2025) <https://www.independent.co.uk>.
[96] Privacy Act 2020 ss 97 and 103.

individual".[97] The HRRT has awarded a wide range of damages for emotional harm, from $10,000 to $98,000.[98] The Tribunal will take into account the severity of harm as well as the conduct of both the individual and the agency. For example, if X Corp removed the deepfake quickly or offered Person A an apology, this may have a mitigating impact on the quantum of damages.[99] Moreover, because X Corp hosts the AI tool (i.e is not the developer) and did not enter the prompt that generated the deepfake, the platform is not the sole cause of the harm. This is likely to also reduce the quantum of damages. The process is also slow. Complaints may take months to be investigated and years to resolve before the HRRT, limiting the remedy's practical effectiveness.[100]

Overall, this pathway illustrates the limits of individualised privacy regulation in addressing AI-enabled harm. GenAI operates through a chain of interconnected decisions that may involve the collection of training data, model development, system integration, prompt processing, content generation, storage, moderation and dissemination. These activities may involve different entities, even though they collectively produce the conditions in which the abuse occurs. That fragmentation risks obscuring the distribution of technological power. Person B may exercise the immediate power to generate the deepfake, but X and xAI exercise forms of institutional and technical power through their control over the system that enables the conduct. The privacy framework may recognise aspects of that relationship, particularly through its technologically neutral definition of personal information, but its enforcement mechanisms are organised around an individual complaint about a particular interference. The result is a regulatory structure that can recognise individual privacy violations without necessarily confronting the technological arrangements through which those violations become foreseeable and repeatable. The burden of navigating regulatory processes therefore falls on those subjected to image-based sexual abuse, while the platform architecture enabling misuse remains structurally the same.

[97] Privacy Act 2020, s 105(1)(d).
[98] Human Rights Review Tribunal "Damages Awarded" *Ministry of Justice* (21 October 2025).
[99] *Hammond v Credit Union Baywide* [2015] NZHRRT 6, [170].
[100] For example, Alison Pugh "Woman's wait for answers three years after human rights hearing" (20 September 2024) *1News* <https://www.1news.co.nz/>.

3 Defamation

Recent developments in defamation law have complicated the question of who "publishes" defamatory material online. At common law, publication traditionally imposed strict liability on any party who intentionally participated in the communication of defamatory content, regardless of knowledge or intent.[101] However, New Zealand courts have adopted a more measured approach. In *Murray v Wishart*, the Court of Appeal held that a Facebook page host is only liable for third-party defamatory content if they have actual knowledge of it and fail to remove it within a reasonable time.[102] This "actual knowledge" test sets a high threshold for liability.

Beyond common law, the HDCA creates a statutory safe harbour for Online Content Hosts, including companies like X Corp. Under ss 23–25, a host is shielded from civil or criminal liability, including defamation, if it did not post the content itself and complies with the notice-and-takedown procedure in s 24. This process requires the host to notify the author of the complaint within 48 hours and to remove the content if the author does not respond within a further 48 hours. The safe harbour protection is procedural rather than automatic. If a complainant does not engage the HDCA process, the safe harbour may not apply, and liability reverts to the general principles of defamation. The existence of the safe harbour raises a broader concern for gendered digital abuse. By insulating online content hosts from liability where the procedural requirements are satisfied, the HDCA places a significant barrier between a person subjected to serious sexualised abuse and any remedy against the platform through which that abuse occurs. In this context, safe harbour can function to shift responsibility away from platforms even where their technologies facilitate foreseeable forms of gendered abuse. The platform may therefore have no legal responsibility for the conditions that enable the abuse, provided it responds once the harm has been reported.

If safe harbour does not apply, a critical issue will be whether the deepfake was "posted" by a user or by X Corp itself. The analysis becomes more complex where the content is generated by xAI's Grok, which is hosted on X. The HDCA safe harbour applies only where the host did not post the material itself or at its direction. If Grok is considered to act on behalf of X Corp, such as where it autonomously generates and publishes a deepfake as part of the platform's infrastructure, X Corp

[101] For example, *Fairfax Media Publications Pty Ltd v Voller* [2021] HCA 27.
[102] *Murray v Wishart* [2014] NZCA 461.

could be deemed the publisher and excluded from safe harbour protection. Conversely, if Grok is treated as a tool through which users generate content, X Corp may retain safe harbour protection, provided it complies with s 24. This characterisation nevertheless risks obscuring the platform's role in designing and deploying a system through which the abuse is enabled.

If X Corp were not protected by the HDCA safe harbour, it could be liable as a publisher under general defamation law. Following *Murray v Wishart*, liability may arise where the platform has actual knowledge of defamatory material and fails to remove it within a reasonable time.[103] Knowledge could arise through direct notification, user reports, media coverage or other circumstances bringing the material to X Corp's attention. In practice, however, a successful claim would face a significant hurdle: unless the deepfake were attributable to X Corp rather than the user, the company would likely remain an Online Content Host and retain the statutory safe harbour if it complied with the HDCA procedure.

4 Negligence

As with defamation, X Corp would likely be treated as an Online Content Host under the HDCA and potentially protected from negligence liability by its safe harbour, provided it complied with the notice-and-takedown regime. Nevertheless, there are additional hurdles. A successful negligence claim requires a duty of care, breach of that duty, causation and actionable damage.

On a duty of care, it is arguably reasonably foreseeable that users of Grok would prompt it to create nonconsensual sexualised deepfakes. Pornographic deepfakes are widespread across the internet, and deepfakes are one of the primary forms of image-based sexual abuse.[104] However, to recognise a duty of care, there must also be sufficient proximity between X Corp and person A.

In traditional cases, proximity is easier to establish where the defendant directly interacts with or undertakes obligations toward the plaintiff (for example, manufacturer-purchaser). In novel contexts the courts may question whether the defendant's deployment of an AI tool creates sufficient relational closeness. While proximity can arise from control over a source of danger,[105] the absence of direct engagement and the action of a third-party user (Person B) complicates the

[103] *Murray v Wishart*, above n 102.
[104] Asher Flynn and others "Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse" (2022) 62 *The British Journal of Criminology* 1622.
[105] *Goldman v Hargrave* [1967] 1 AC 645.

analysis. The inquiry becomes whether the platform's architecture and foreseeable misuse establish a relationship akin to assumption of responsibility, or whether the harm is too attenuated to justify a duty. Courts are ordinarily reluctant to find proximity where recognising a duty would expose a defendant to an indeterminate number of claims.[106] However, this objection arguably weakens where the technology is designed or deployed in a way that directly exposes identifiable individuals to a specific risk. In the context of sexualised deepfakes, the affected class is arguably more targeted than in the general run of online harm cases given the overwhelming use of image models to target women and girls. The distribution of this risk is relevant to foreseeability, namely, where a technology is repeatedly used to sexualise women without their consent, the resulting harm is a predictable consequence of its operation rather than an exceptional misuse by an isolated user. We argue that a foreseeability analysis that treats each incident as an individual act by Person B risks overlooking the gendered pattern of harm that makes the risk foreseeable in the first place.

A second complexity is that the claim asks the court to recognise a positive duty to prevent harm caused by third-party misuse of an AI system. New Zealand courts are generally cautious with omissions, especially where the defendant has not assumed responsibility for the particular plaintiff and where the immediate harm is inflicted by a third party.[107] Nevertheless, the law does recognise positive duties where the defendant creates or controls a source of danger[108] or the "sparking of danger" line where the defendant enables foreseeable wrongdoing by others by omission to deal with a special danger in which they have control.[109] In analogy, one could argue that X Corp integrated Grok which presented a foreseeable danger, the risk it would be used by third parties to create nonconsensual deepfakes. Many cases argued along these lines have failed where there is insufficient proximity.[110]

A systems-based approach to foreseeability may support recognising a duty of care in this context. Fraser and Suzor argue that the relevant question is not whether an individual user misused an AI tool, but whether the harm was foreseeable from the way the system was designed, deployed and

[106] *Van Der Wetering v Capital Coast Health Ltd* HC Wellington CP368/98, 19 May 2000.
[107] *North Shore City Council v Attorney-General (The Grange)* [2012] 3 NZLR 341.
[108] The occupier who must abate a hazard thrust upon them see *Goldman v Hargrave*, above n 105.
[109] *Topp v London Country Bus (South West) Ltd* [1993] 1 WLR 976. See also *Smith v Littlewoods Organisation Ltd* [1987] AC 241 wherein the defendants owned an empty building that was set on fire by vandals causing damage to neighbouring properties. Lord Goff opined that where the defendant knew or had the means to know that third parties have created or are creating a risk of danger then they had a duty to prevent the damage to neighbouring properties.
[110] For example, *Smith v Littlewoods Organisation Ltd,* above n 109.

governed.[111] Risk is generated upstream through the designer's/developer's choices about training, safeguards and moderation, and may include predictable patterns of misuse. Where X Corp controls Grok's architecture, and derives commercial benefit from its deployment, it is arguable that harm resulting from its operation should not be treated as too remote simply because a user supplied the prompt.

However, there are significant limitations in negligence as a regulatory response to AI-mediated harm.[112] Negligence remains oriented toward discrete actors, identifiable moments of fault, and linear causal chains, whereas GenAI harm is distributed across multiple actors and decisions within a complex socio-technical system. This makes it difficult to localise breach or causal responsibility in a way that maps cleanly onto traditional doctrinal analysis. As a result, negligence risks under-capturing systemic risk while simultaneously over-focusing on *ex-post* harm (addressing harm after the fact) rather than *ex-ante* design and governance choices (addressing risk of harm occurring).

On breach, the core question is whether X Corp took reasonable care to prevent a type of harm that was both foreseeable and avoidable. New Zealand courts typically look at several practical factors when deciding this, namely, how likely the harm was, how serious the consequences could be, how useful the activity is, and how easy it would have been to take steps to reduce the risk.[113] Sexualised deepfakes are known to cause serious psychological harm,[114] and basic safeguards such as prompt blocking, output filtering, and deepfake-detection tools are technically feasible. On that basis, a court could find that failing to implement these measures fell below reasonable care, particularly given how widespread deepfake abuse has become. The relevant question therefore extends beyond whether precautions were technically available to whether X took reasonable steps to protect a group that was predictably exposed to this particular form of abuse.

There are, however, meaningful counterarguments. X Corp may argue that the suggested precautions are not as simple or reliable as portrayed: large-scale filtering and deepfake detection remain technically challenging, and overly strict guardrails risk blocking legitimate uses of the

[111] Henry Fraser and Nicolas Suzor "Locating fault for AI harms: a systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain" [2025] 17 *Law, Innovation and Technology* 103.
[112] Fraser and Suzor, above n 111, at 111-115.
[113] *Wyong Shire Council v Shirt* (1980) 146 CLR 40; *Wilson & Horton Ltd v Attorney-General* [1997] 2 NZLR 513.
[114] Alexander Diel and others "The harm of deepfakes: a scoping review of deepfakes' negative effects on human mind and behavior" (2025) *AI & Society* 1.

tool. The company could also contend that the immediate cause of harm was Person B's deliberate misuse, not the AI tool's design framing the issue as a failure to stop another person's wrongdoing, something courts are generally cautious about treating as negligence. Finally, X Corp might argue that decisions about AI safety involve broader policy choices better suited to Parliament than to case-by-case judicial rulings.

The decisive constraint is damage: negligence requires a recognised psychiatric injury, rather than serious emotional distress or humiliation alone.[115] The damage requirement operates as a substantive limit on the law's capacity to respond to nonconsensual sexualised deepfakes. The primary harms associated with this material (sexual humiliation, loss of bodily and sexual autonomy, and the ongoing fear of further circulation) are well documented, however, remain legally unrecognised unless they manifest as a clinically diagnosed psychiatric disorder. This reflects a broader tendency within negligence law to discount harms that are gendered and sexualised unless they can be translated into medicalised injury.[116] It also illustrates the difficulty of applying an individualised damage model to harm produced through technological systems, where the loss of control over sexual representation may be the central injury even without a clinically recognised condition. In practical terms, this constrains the usefulness of negligence as a response to AI-facilitated sexual abuse: even where foreseeability and breach are arguable, most claims will fail at the damage stage.

Part IV has highlighted the limitations of New Zealand's existing legal and regulatory framework in holding platforms accountable for AI-enabled sexualised deepfakes. The available avenues offer uncertain, largely retrospective remedies, while imposing few enforceable obligations on platforms to assess and mitigate the foreseeable gendered harms arising from the systems they develop or deploy.

V Conclusion

This article has examined the capacity of New Zealand law to respond to AI-enabled gendered abuse, using the integration of GenAI into social media platforms and nonconsensual sexualised

[115] *Van Soest v Residual Health Management Unit* [2000] 1 NZLR 179.
[116] For example, Martha Chamallas "The Architecture of Bias: Deep Structures in Tort Law" (1998) 146 *University of Pennsylvania Law Review* 463; Yifat Bitton "Liability of Bias: A Comparative Study of Gender" (2010) 16 *Annual Survey of International and Comparative Law* 63.

deepfakes as a case study. The analysis demonstrates a recurring mismatch between existing legal frameworks and the conditions in which this abuse occurs. Criminal and civil mechanisms provide some avenues for redress against an individual perpetrator, but they tend to organise responsibility around a narrowly defined act, an identifiable malicious actor and proof of particular forms of legally recognised injury. Sexualised deepfakes arise within a technological environment shaped by decisions about training data, system design, platform integration and safeguards. The legal focus on the individual user therefore captures only part of the conditions through which the abuse occurs.

This is particularly significant in the context of gendered abuse. Sexualised deepfakes disproportionately target women and reproduce longstanding forms of sexualisation and control over women's bodies through technological systems. However, the existing legal framework repeatedly asks the person subjected to that abuse to establish a form of harm that fits an existing legal category. Section 22 of the HDCA requires proof of serious emotional distress and specific intent, while negligence requires recognised psychiatric injury. Other causes of action struggle with the synthetic nature of the image or the role of the platform through which it was generated and disseminated. The result is that the law can recognise aspects of the conduct while leaving the underlying loss of sexual autonomy only partially addressed.

The proposed Deepfake Digital Harm and Exploitation Bill would improve this position by bringing synthetic images within the statutory framework governing intimate visual recordings. It would provide a clearer basis for addressing the creation and dissemination of sexualised deepfakes and reduce the distinction currently drawn between genuine and synthetic images. The Bill nevertheless addresses only part of the wider problem. Its focus remains on particular acts involving deepfake content, leaving other forms of harm and points in the technological process outside its reach.

The analysis of platform responsibility exposes a further limitation. X and xAI occupy different positions within the technological structure, but both exercise significant control over the conditions in which sexualised deepfakes can be produced and circulated. Existing law has difficulty attaching responsibility to those upstream decisions. The voluntary nature of the current online safety framework, the protection provided by safe harbour, the fragmented application of privacy law and the limits of negligence each create different routes through which platforms can

distance themselves from foreseeable misuse. These difficulties are not unique to X or Grok, but reflect a broader problem in regulating GenAI systems capable of facilitating gendered abuse that are integrated into existing social networking platforms or websites.

The central issue for future regulation is where responsibility should sit when a technological system creates a foreseeable capacity for gendered abuse. A framework centred principally on the individual who enters the prompt responds after the technological conditions for abuse has already been created. For the framework to be truly effective, however, greater responsibility should be placed on those who design, develop and deploy systems with the knowledge of foreseeable misuse. One such avenue could be drawn from principles of product liability, under which producers are required to take reasonable steps to ensure the safety of their products, with those who fail to do so potentially facing significant financial penalties or restrictions on their ability to operate. Such an approach would represent a shift from a predominantly reactive framework to one more focused on the prevention and mitigation of non-consensual sexualised deepfakes. It is an unfortunate reality that change of this kind requires political will, particularly where regulation may be met with resistance from powerful technology companies. As such, any undertaking of this kind will depend on the Government of the day being willing to impose meaningful obligations on companies whose products and services contribute to the conditions in which this form of abuse occurs.

The case study ultimately demonstrates that addressing AI-enabled gendered harm requires attention to both the immediate perpetrator and the technological environment in which the conduct occurs. Therefore, the challenge for New Zealand law is to develop forms of accountability capable of responding to harm that is enabled through technological systems and disproportionately directed at women. Any proposed reforms should ensure that responsibility follows the capacity to create and reduce the risk, rather than resting primarily with those subjected to the resulting abuse.