Ambient Non-Consensual Image Synthesis: A Convergence Threat Analysis of AR Wearables, Real-Time Video Generation, and Open-Source Nudification Models

Travis Gilly Real Safety AI Foundation

Abstract

We identify and analyze an emergent threat arising from the convergence of four maturing technologies: (1) open-source AI nudification models [1], (2) real-time video generation systems [2][3][4], (3) consumer AR wearables with camera input [5][6], and (4) edge and cloud AI inference [7][8]. While each component has received independent scholarly attention, no prior work synthesizes these into a unified threat model [9][10][11]. We demonstrate that real-time ambient non-consensual intimate image synthesis is architecturally feasible today via cloud-assisted pipelines [2][3], and will be feasible on-device within 12 to 24 months given current diffusion acceleration and edge hardware trajectories [12][13]. We further show that proposed mitigations (on-device safety filters) are fundamentally vulnerable to adversarial bypass, as illustrated by attacks on Google's SafetyCore system [14][15]. We introduce the concept of "perceptual consent" to describe the novel harm category where individuals lose autonomy over how their bodies are rendered in others' visual fields, grounding this concept in established frameworks of bodily autonomy, informational self-determination, and human dignity [16][17][18]. We document that population-level psychological harm from the technology's mere existence is already empirically established in the social media deepfake context and will intensify as the threat extends from digital to physical space. We conclude with technical, policy, and research recommendations.

---

1. Introduction

1.1 The Synthesis Gap

The past three years have seen rapid maturation across several technology domains that, while developed independently, share a troubling potential for convergence. AI-powered nudification tools have proliferated across web platforms, with systematic

analyses identifying more than twenty commercial applications and thousands of open-source models [1]. Real-time video generation has progressed from minute-long rendering times to sub-second latency in cloud environments [2][3][4]. Consumer AR wearables with integrated cameras have reached mass-market availability [5][6]. Edge AI inference has achieved the performance necessary for real-time computer vision on mobile and wearable hardware [7][8].

Each of these domains has attracted scholarly attention. Researchers have examined the ecosystem of nudification platforms and their harms [1][19][20]. Privacy scholars have analyzed data collection practices in VR and AR environments [9][21][22]. Computer vision and machine learning researchers have benchmarked video generation systems and edge inference capabilities [12][13][7][8]. Yet these research communities have operated largely in isolation from one another.

The result is a critical synthesis gap. No prior work examines what happens when these technologies converge: when someone wearing commercially available AR glasses can, in real time, render synthetic nude versions of fully clothed people in their field of view. This paper addresses that gap.

The urgency of this analysis is underscored by recent events. In January 2026, the AI assistant Grok, integrated into the social platform X, was used to generate hundreds of thousands of non-consensual intimate images of real people, including minors [23][24][25]. The incident demonstrated that demand for automated non-consensual undressing is not hypothetical; it is immediate, coordinated, and operates at scale when tools become accessible. The California Attorney General launched a formal investigation [24]. Yet even this incident was discussed in isolation from the wearable and real-time rendering technologies that would make such abuse ambient and continuous rather than discrete and platform-mediated.

1.2 Contribution

This paper makes six contributions to the literature on AI safety, privacy, and synthetic media harms.

First, we provide the first unified threat analysis that synthesizes AR wearables, nudification models, and real-time video generation into a coherent threat model, structured according to established threat modeling frameworks including STRIDE [26]. We demonstrate that the architectural components necessary for ambient non-consensual nudity rendering are either already available or on clear development trajectories.

Second, we establish a technical feasibility timeline with concrete latency benchmarks drawn from current systems including TurboDiffusion, SnapGen-V, and PixVerse R1 [12][13][2][3]. We provide sensitivity analysis across conservative, moderate, and

aggressive scenarios to acknowledge uncertainty while demonstrating that directional conclusions are robust.

Third, we introduce the concept of "perceptual consent" as a formal framework for understanding a novel category of harm, grounding this concept in established philosophical frameworks including Martha Nussbaum's capability approach [16], Daniel Solove's privacy taxonomy [17], and Kantian conceptions of human dignity [18]. We demonstrate that traditional consent frameworks are structurally inapplicable to ambient synthesis and establish principled criteria distinguishing rights-violating synthesis from mere imagination.

Fourth, we document population-level psychological harm caused by the technology's existence independent of individual victimization, synthesizing empirical research on chilling effects, anticipatory threat perception, and collective cultural injury [71][72][77][78][79].

Fifth, we demonstrate that the most commonly proposed mitigation (on-device safety filters) is fundamentally insufficient, drawing on seminal work in adversarial machine learning [27][28][29] and recent research showing successful attacks against Google's SafetyCore system [14][15].

Sixth, we provide a policy and research agenda targeting this specific convergence, identifying gaps in current regulatory frameworks including the EU AI Act [30], US federal legislation including the DEFIANCE Act [31] and TAKE IT DOWN Act [32], and UK law including the Online Safety Act 2023 [33].

1.3 Scope and Limitations

This analysis focuses specifically on synthetic non-consensual explicit and abusive content imagery (SNEACI), rather than the broader category of deepfakes or synthetic media [1][19]. While political deepfakes and other synthetic media raise significant concerns, the specific harm profile of non-consensual intimate imagery warrants focused analysis. Prevalence studies indicate that approximately 1 in 12 adults have experienced image-based sexual abuse, with AI-generated imagery representing a growing proportion [34][35].

We further focus on ambient and real-time rendering in AR contexts rather than static image generation or offline video deepfakes [2][13]. The shift from discrete, platform-mediated content creation to continuous, perception-layer augmentation represents a qualitative change in the nature of the threat.

We acknowledge uncertainty in timeline projections. Section 3.5 provides sensitivity analysis across three scenarios. Hardware availability, model optimization, and deployment choices by platform operators could accelerate or delay the scenarios we

describe. However, the direction of development is clear, and the architectural feasibility we establish does not depend on precise timeline accuracy.

1.4 Dual-Use Considerations

We acknowledge that the component technologies we examine have legitimate applications. AR wearables enable accessibility tools, navigation assistance, and professional applications [36]. Real-time video generation enables creative expression and entertainment [2]. Edge AI enables privacy-preserving computation [7].

This dual-use character is common to many technologies that have been subject to conduct-specific regulation. Three-dimensional printing enables both prosthetics and untraceable weapons; regulation targets the weapons, not the printers. Cryptography enables both privacy and criminal coordination; regulation targets specific criminal uses, not encryption itself. We argue for similar conduct-specific approaches to ambient non-consensual synthesis, targeting the harmful application rather than the underlying technologies.

1.5 The Escalation of Ambient Threat

The threat we describe is not novel in kind but in degree. Non-consensual sexualization of women's bodies has existed across technological eras; what changes is the scope of the "safe zone" within which women can exist without rational fear of such violation. We identify four stages of escalation, each enabled by technological shifts:

Stage 1 (Pre-digital): Women controlled exposure through physical proximity. Non-consensual sexualization required physical access to the person or to physical photographs. The safe zone encompassed all spaces where one was not physically observed.

Stage 2 (Early internet): Women controlled what they shared digitally. Risk was bounded by voluntary disclosure. A woman who did not post photographs online faced minimal risk of digital sexualization. The safe zone contracted to exclude voluntarily shared digital spaces but preserved physical and unshared digital environments.

Stage 3 (Deepfake/AI nudification era): Any shared image becomes source material for synthesis. The January 2026 Grok incident demonstrated that 102 "undressing" requests could occur in a ten-minute span, targeting unknown women from publicly available photographs [23][24][25]. The safe zone contracted further: women now face rational fear when posting any photograph, regardless of content or context. Research documents consequent chilling effects, with women withdrawing from online participation to avoid potential targeting [71][72].

Stage 4 (Ambient AR synthesis): Physical presence in public space becomes sufficient for targeting. No photograph need be shared; existing in view of AR-equipped observers provides source material for real-time synthesis. The safe zone contracts to private spaces where no observer with AR capability is present, effectively eliminating safe participation in public life.

Each stage represents not merely quantitative expansion but qualitative transformation of the threat surface. Stage 3 harms are now empirically documented: self-censorship, withdrawal from platforms, anticipatory anxiety affecting women who have never been individually targeted [71][73][74]. Stage 4 extends these documented harms from digital to physical space, from asynchronous capture to real-time perception, from image-based to existence-based vulnerability.

The progression reveals a consistent pattern: technological capability expands, the burden of avoidance shifts to potential victims, and the domain of unencumbered participation shrinks. Our analysis of Stage 4 builds on the empirical foundation established for Stage 3 harms.

1.6 Responsible Disclosure Considerations

We acknowledge that publishing this threat analysis may itself contribute to anticipatory harm. By documenting the feasibility and timeline of ambient non-consensual synthesis, we risk accelerating public awareness of a threat before it materializes at scale, potentially producing the population-level psychological harm we describe in Section 4.7.

We have weighed this risk against the alternative: allowing the threat to materialize without advance warning to policymakers, platforms, researchers, and civil society. The component technologies we describe are public: PixVerse R1's real-time capabilities are marketed openly [2][3], nudification models are freely distributed on model-hosting platforms [1][37][38], and consumer AR glasses are available in retail stores [5][6]. Anyone attending to these domains can identify the convergence trajectory independently.

The question is not whether this convergence will be recognized but who recognizes it first and what they do with that recognition. If researchers remain silent, bad actors build quietly and harm accumulates before defensive frameworks exist. If researchers publish, harm from awareness occurs, but policymakers and civil society gain lead time for response.

We have attempted to balance warning against weaponization by focusing on threat characterization rather than implementation guidance, providing policy and research recommendations rather than attack tutorials, targeting academic and policy

audiences through venue selection, and pre-briefing relevant stakeholders where possible.

This threat analysis is intended to enable defense, not to enable attack. We recognize that intention does not guarantee outcome and that some readers may misuse our analysis. We judge that the benefits of advance warning outweigh the costs of premature awareness, but we acknowledge reasonable disagreement on this balance.

---

2. Background and Related Work

2.1 AI Nudification Ecosystem

The ecosystem of AI-powered tools for generating non-consensual intimate imagery has grown rapidly since the introduction of accessible diffusion models in 2022. Gibson et al. provide the most comprehensive systematic analysis to date, examining more than twenty commercial nudification web applications [1]. Their analysis documents the features, monetization models, and targeting patterns of these platforms, finding that the vast majority are designed to generate non-consensual imagery of women without meaningful age verification or consent mechanisms.

The scale of the ecosystem extends beyond commercial platforms. Hugging Face and CivitAI host thousands of models trained for undressing and sexualization purposes, including explicitly "uncensored" model collections designed to bypass platform safety restrictions [37][38]. The University of Florida research team estimates that per-image generation costs range from zero to six cents, and emphasizes that "anybody can do this" via anonymous web interfaces with no technical barriers to entry [39].

The normalization of these tools has been documented in qualitative research. A 2025 study examining user behavior on nudification platforms found that users frequently justify their actions through claims of harmlessness, framing synthetic imagery as victimless because no "real" exposure occurred [20]. This rationalization pattern suggests that ambient AR-based nudification would face similar normalization pressures.

While web-based platforms dominate, early hardware prototypes have attempted to bridge the gap between capture and synthesis. The "NUCA" camera project (Vef & Groß, 2024), introduced as an art intervention, demonstrates a standalone device capable of generating nude imagery from camera input [82]. However, NUCA relies on cloud-based inference with latencies exceeding 10 seconds, functioning as an automated deepfake camera rather than an augmented reality device. It illustrates the conceptual interest in capture-to-synthesis pipelines but fails to achieve the real-

time, perception-layer integration that characterizes the ambient threat described in this paper.

The January 2026 Grok incident provides the most recent illustration of demand [23][25][24]. Within days of users discovering that Grok would generate non-consensual intimate imagery on request, coordinated communities on X used the tool to undress hundreds of thousands of real people, including public figures and minors. Platform restrictions were implemented only after significant public pressure and regulatory attention.

2.2 Real-Time Video Generation

Video generation using diffusion models has progressed rapidly from research demonstrations to commercial products. Early systems exhibited per-clip generation times on the order of minutes, making real-time or interactive applications infeasible. By late 2024, production tools including Runway, Pika, Kling, and PixVerse V5 had reduced generation times to 10 to 60 seconds for 8 to 10 second clips at 1080p resolution [40][41][42].

Two recent developments suggest that the latency barrier to real-time generation is falling rapidly.

TurboDiffusion, announced in December 2025, reports 100 to 200 times speedup over standard diffusion sampling through sparse attention mechanisms and rectified consistency models [12]. The authors demonstrate 8-second video generation in under 8 seconds on high-end GPU hardware, approaching the real-time threshold.

SnapGen-V demonstrates that mobile devices can approach real-time video generation through aggressive model distillation [13]. The system generates "five-second video within five seconds on a mobile device" using one-step latent diffusion, indicating that edge devices are approaching practical clip generation without cloud assistance.

Most significantly, PixVerse R1, released in January 2026, is presented as a real-time world model [2][3][4][43]. Rather than generating fixed-length clips, R1 uses an "Instantaneous Response Engine" that reduces diffusion sampling to 1 to 4 steps, enabling interactive 1080p video that "responds instantly to user commands" and supports infinite-length streaming.

2.3 AR Wearables

Consumer AR wearables with integrated cameras have reached mass-market availability. Meta's Ray-Ban smart glasses integrate 12 megapixel cameras, microphones, and an AI assistant into a form factor resembling conventional eyewear

[5][6]. The glasses can continuously capture video, and reporting has documented both privacy concerns and instances of covert recording [5].

Privacy analyses highlight that data from Ray-Ban glasses is processed both on-device and in Meta's cloud infrastructure, with limited transparency regarding data retention and consent mechanisms for bystanders [6][44]. The LED indicator intended to signal recording can be obscured or disabled [5].

AR privacy research has examined data collection concerns but has not addressed real-time synthetic body modification. Speicher et al. [45] and Williams et al. [46] examine AR privacy through the lens of data flows and bystander awareness, but assume the captured imagery represents reality rather than a substrate for synthetic transformation.

2.4 Edge AI and On-Device Inference

Edge AI has matured to the point where real-time computer vision on wearable-class devices is routine [7][8][47]. Wearable and AR research prototypes demonstrate that non-trivial AI workloads can run entirely on glasses-class hardware, including real-time first-aid assistance [36] and vision assistance for the visually impaired [48]. Performance evaluations of current XR devices confirm that local inference for interactive workloads is achievable within acceptable latency and power envelopes [49].

2.5 Adversarial Machine Learning

The vulnerability of neural networks to adversarial examples has been established since Goodfellow et al.'s seminal 2014 work demonstrating that imperceptible perturbations can cause arbitrary misclassification [27]. Subsequent work by Carlini and Wagner [28] and Papernot et al. [29] demonstrated that adversarial examples are robust across different attack scenarios and that defenses are consistently bypassable.

This body of work is directly relevant to proposed on-device safety filters. Any neural network deployed for content filtering can be attacked using established techniques, as demonstrated in the SafetyCore case [14][15].

2.6 Identified Gap

Our review reveals a consistent pattern: each component technology has received scholarly attention, but no work synthesizes them into a unified threat model for ambient non-consensual nudity rendering.

VR and AR privacy research emphasizes tracking, biometric inference, and behavioral profiling, but does not treat real-time synthetic modification of bystanders' bodies as a distinct risk [21][22][50]. Synthetic media research focuses on content authenticity

and distributed media harms [19][51][52]. Nudification research focuses on web platforms [1][39][20]. AI safety indices do not mention AR-based non-consensual nudity rendering [53].

This paper addresses this synthesis gap directly.

---

3. Threat Model

3.1 Attack Scenario

We describe a threat scenario using the STRIDE framework [26] to systematically identify threat categories.

Scenario: An attacker uses commercially available or near-term AR glasses to render synthetic nude versions of fully clothed people in their field of view, in real time, without the knowledge or consent of those people.

The attack proceeds as follows:

1. The attacker wears AR glasses equipped with cameras and heads-up displays, combining the form factor of current smart glasses (e.g., Meta Ray-Ban) with the display capabilities of emerging prototypes (e.g., Meta Orion) [5][6].
2. The glasses continuously capture video of surroundings, including fully clothed bystanders.
3. Edge processing performs person detection, tracking, and pose estimation using standard computer vision pipelines [7][8].
4. Either locally or via cloud, undressing models process segmented body regions, generating synthetic nude versions aligned to current pose [1][2][13].
5. Synthetic output is composited onto the AR display, visible only to the attacker.
6. Optionally, the attacker records or streams the synthetic output.

STRIDE Analysis:

- **Spoofing**: The system spoofs the victim's appearance in the attacker's perception
- **Tampering**: The visual representation of the victim is tampered with without authorization
- **Repudiation**: No evidence need exist; the attack is deniable

- **Information Disclosure**: The victim's imagined nude body is "disclosed" to the attacker
- **Denial of Service**: N/A
- **Elevation of Privilege**: The attacker gains unauthorized perceptual access to the victim's body

3.2 Technical Pipeline

The attack pipeline can be decomposed into discrete processing stages:

Stage	Latency	Notes
Camera input	<1ms	Continuous 30fps capture
Person detection + tracking	15-30ms	YOLOv5 or similar on edge hardware [7][8]
Region crop + pose estimation	20-30ms	Standard pose models
Cloud upload (if applicable)	50-100ms+	Network dependent
Undressing model inference	100-500ms	Model and hardware dependent [12][54]
Video synthesis/interpolation	100ms-5s	Architecture dependent [12][13][2]
AR display composition	5-10ms	Standard rendering

Total end-to-end latency ranges from approximately 200ms (cloud-assisted, fast models) to several seconds (edge-only, full video synthesis).

We note that thermal throttling on wearable form factors may reduce sustained performance under continuous operation. The likely near-term attack vector is not standalone glasses but glasses tethered to a smartphone via WiFi Direct or UWB, with the phone serving as the compute platform. This tethered configuration resolves thermal and battery constraints while maintaining the ambient, hands-free nature of the threat. Even for standalone operation, intermittent synthesis patterns consistent with sustained mobile gaming workloads remain feasible.

3.3 Feasibility Classes

Class	Latency	Feasibility	Quality	Requirements
Cloud-assisted (PixVerse-class)	200-500ms	NOW	High	Network connectivity
Edge static + interpolation	200-400ms	NOW	Medium	Current mobile hardware
Edge full video synthesis	2-5s	12 months	High	Next-gen edge chips
Edge real-time (<500ms)	<500ms	24-36 months	High	Future optimization

We note that the full glasses-based attack (capture and display in single device) awaits convergence of camera and display in discreet form factors such as Meta Orion prototypes. However, a variant attack is feasible today: capture via Ray-Ban Meta glasses, stream to phone, cloud synthesis, view on phone screen. This variant is less immersive but achieves the same ambient, covert, real-time synthesis.

3.4 Threat Actors

We identify four threat actor categories:

1. **Individual harassers**: The documented user base of nudification platforms skews toward individuals targeting specific known victims [55][20].
2. **Coordinated communities**: The Grok incident demonstrated rapid community formation around accessible tools [23][25].
3. **Commercial exploiters**: Subscription services offering nudification on demand could extend to AR-based offerings [1][39].
4. **State actors**: Documented interest in synthetic media for kompromat and psychological operations [19][56].

3.5 Timeline Sensitivity Analysis

Given uncertainty in technological development, we provide three scenarios:

Conservative Scenario (assumes slower optimization, hardware delays):

- Cloud-assisted real-time: Available now
- Edge static + interpolation: Available now
- Edge full video synthesis: 18-24 months
- Edge real-time: 36-48 months

Moderate Scenario (assumes current trajectory continues):

- Cloud-assisted real-time: Available now
- Edge static + interpolation: Available now
- Edge full video synthesis: 12 months (Medium/High quality)
- Edge real-time: 24-36 months

Aggressive Scenario (assumes breakthrough optimization, existing silicon pushed beyond nominal specs):

- Cloud-assisted real-time: Available now
- Edge static + interpolation: Available now
- Edge full video synthesis: 9-12 months
- Edge real-time: 18-24 months

Key dependencies:

- One-step diffusion optimization (TurboDiffusion trajectory) [12]
- Mobile NPU advancement (Snapdragon, Tensor roadmaps)
- Model quantization and distillation research [13][54]
- World model architectures (PixVerse R1 trajectory) [2]

In all scenarios, cloud-assisted real-time rendering is feasible today. The question is not whether the threat will materialize, but when edge-only deployment becomes practical.

---

4. The Perceptual Consent Problem

4.1 Philosophical Grounding

Before introducing perceptual consent as a concept, we ground it in established philosophical frameworks.

Bodily Autonomy and the Capability Approach

Martha Nussbaum's capability approach identifies bodily integrity as a central human capability: "Being able to move freely from place to place; to be secure against violent assault... having opportunities for sexual satisfaction and for choice in matters of reproduction" [16]. While Nussbaum's framework addresses physical bodily integrity,

we argue it extends to representational bodily integrity, the ability to control how one's body is represented and perceived by others.

The capability approach asks not merely whether a harm has occurred, but whether an individual's capability to function with dignity has been diminished. Ambient non-consensual synthesis diminishes the capability to exist in shared physical space without one's body being rendered in degrading ways by others.

Informational Self-Determination

Daniel Solove's taxonomy of privacy harms provides a framework for understanding information-related injuries [17]. Solove identifies "exposure," defined as revealing another's nudity or intimate activities, as a distinct privacy harm. He also identifies "increased accessibility," making information more easily obtainable.

Ambient synthesis creates a novel variant: the victim's actual body is not exposed, but a synthetic representation is generated and "exposed" to the attacker. The harm resembles exposure without the predicate of actual nudity. Solove's framework suggests this constitutes a privacy harm even absent distribution, as the violation occurs in the generation itself.

Dignity and the Kantian Framework

Kantian ethics holds that persons must be treated as ends in themselves, never merely as means [18]. Using another person's body as raw material for sexual imagery, without their knowledge or consent, treats them as a mere means to the attacker's ends.

This dignity violation does not require that the victim know about it. A person whose image is used for non-consensual synthesis has been treated as a mere means regardless of their awareness. The harm is to their status as a person deserving of respect, not merely to their psychological state.

4.2 Defining Perceptual Consent

Building on these philosophical foundations, we introduce "perceptual consent" to name what is violated in ambient non-consensual synthesis.

Definition: Perceptual consent refers to an individual's autonomy over how their body is rendered in others' visual perception. When someone synthesizes intimate imagery of another person, overlaying it on their perception, they violate that person's perceptual consent regardless of whether any artifact is created, stored, or shared.

This concept extends prior work on consent in immersive environments. Research on VR consent has documented how immersive systems collect data that users cannot meaningfully understand or consent to [9][57]. Work on spontaneous AR interactions

has explored how consent mechanisms fail when interactions are ambient [58]. Work on AR bystander privacy has shown that individuals struggle to control how they are captured by others' devices [10].

We extend these analyses from data collection to perceptual manipulation. The harm is not merely that data is collected, but that one's body is rendered in a degrading manner in another person's experience.

4.3 Comparison to Existing Legal Frameworks

Perceptual consent violations fall outside existing legal frameworks. We analyze four categories of potentially applicable law, demonstrating that each fails to cover the threat.

Voyeurism Laws

Voyeurism statutes typically require observation of private acts or private body parts. The UK Voyeurism (Offences) Act 2019, Section 67A, specifically addresses "operating equipment beneath the clothing of another person" to observe or record genitals, buttocks, or underwear [59].

The statutory language reveals the gap:

"A person (A) commits an offence if... A operates equipment beneath the clothing of another person (B)... with the intention of enabling A or another person (C), for a purpose mentioned in subsection (3), to observe... (i) B's genitals or buttocks... or (ii) underwear covering B's genitals or buttocks, in circumstances where the genitals, buttocks or underwear would not otherwise be visible."

This language targets observation of actual body parts that are concealed. Ambient synthesis captures the fully clothed body and synthesizes nudity; no actual body part is observed. The statutory requirement of operating equipment "beneath the clothing" is not met.

By the interpretive principle of *expressio unius est exclusio alterius* (the expression of one thing excludes others), the explicit focus on equipment beneath clothing suggests the legislature did not intend to cover synthesis of nudity from clothed imagery.

Non-Consensual Intimate Imagery (Revenge Porn) Laws

US state laws on non-consensual intimate imagery typically require distribution. California Penal Code § 647(j)(4) criminalizes distribution of intimate images without consent where the person "knew or should have known" the image was intended to remain private [60]. California Civil Code § 1708.85 provides civil remedies for distribution [61].

Virginia's statute (Code § 18.2-386.2), one of the earliest and broadest, explicitly covers AI-generated imagery but still requires that images be "sold, given, distributed, or published" [62].

Texas Penal Code § 21.165 specifically addresses deepfakes but requires that the imagery be created "without the effective consent" and with intent to harm, with remedies focused on distributed content [63].

The common thread: existing statutes assume distribution as the vector of harm. Perception-only synthesis, where the attacker views but does not share, falls outside their scope.

Harassment Laws

Harassment laws typically require communication to the victim or conduct the victim can perceive. Ambient synthesis involves no communication; the victim does not know it is occurring. Without awareness, harassment frameworks do not apply.

CSAM Laws

Child sexual abuse material laws focus on the existence of material depicting abuse. If no image is saved (perception-only), the legal status becomes uncertain. While some jurisdictions have expanded CSAM definitions to include AI-generated imagery, enforcement assumes artifacts that can be discovered.

4.4 The Rendering Threshold: Distinguishing Imagination from Synthesis

A potential objection to perceptual consent as a rights-bearing concept is that it appears to criminalize imagination. If someone imagines another person nude, they engage in mental activity that, however objectionable, has never been subject to legal prohibition. Why should AI-assisted synthesis differ?

We identify three properties that distinguish silicon rendering from neural imagination, establishing a principled threshold for rights violation.

Realism and Functional Equivalence

Human imagination produces subjective, typically imprecise mental imagery. The imaginer does not know what the target's body actually looks like; they are constructing a fantasy bearing uncertain resemblance to reality. AI synthesis, by contrast, attempts photorealistic output calibrated to the target's actual appearance, body proportions, and current pose. The psychological experience of viewing AI synthesis is closer to viewing actual footage than to daydreaming.

This distinction parallels why deepfakes cause harms that erotic fiction does not. Both are "fake" representations of real people. But deepfakes' realism makes them

functionally equivalent to genuine intimate imagery in ways that prose or crude drawings never achieve. The violation occurs when the representation achieves sufficient realism that it functions as genuine intimate imagery in the viewer's experience.

Externalization to Artifact

Imagination is purely internal; no artifact exists outside the imaginer's mind. Synthesis externalizes the representation onto silicon, creating a computational artifact, even if transiently stored in RAM and never saved to persistent storage. This externalization is the hook for legal and ethical analysis: the artifact exists in the world, can be observed, and could in principle be recorded, shared, or discovered.

The law routinely distinguishes thought from action, intention from execution. Imagining a crime is not criminal; taking steps toward execution may be. Synthesis constitutes execution: computational resources are deployed, data is processed, an output is rendered. The artifact's transience does not erase its existence.

The wiretapping analogy is instructive. In wiretapping law, the harm is the interception itself; recording is secondary. A wiretap that is never recorded still violates privacy because the communication was accessed without authorization. Perceptual synthesis is "visual wiretapping" of the body: the harm is the unauthorized rendering, regardless of whether the output persists. This framing grounds the "externalization" criterion in established legal precedent while clarifying that transient artifacts are sufficient for rights violation.

Accuracy and Specificity

Imagination applies to generic or idealized bodies; the imaginer cannot accurately represent specific anatomical details they have not observed. AI synthesis trained on large datasets produces anatomically plausible outputs specific to the target's observable characteristics. The synthesis is not generic fantasy but targeted representation.

These three properties, realism sufficient for functional equivalence, externalization to computational artifact, and targeted specificity, establish the threshold at which perceptual consent is violated. Mental activity below this threshold, however distasteful, does not cross into rights violation. Synthesis above this threshold does.

4.5 Typology of Statutory Gaps

We identify four categories of gaps in current legal frameworks:

Gap Type	Description	Examples
Creation vs. Distribution	Laws require distribution; creation alone is not prohibited	CA Penal Code § 647(j)(4); VA Code § 18.2-386.2
Real vs. Synthetic	Laws require "real" imagery or body parts	UK Voyeurism Act "beneath clothing" language
Perception-Only Harm	Laws assume artifacts exist; perception-only violations not covered	All surveyed jurisdictions
Wearable AI Disclosure	No requirements to disclose AI capabilities on AR devices	No applicable law identified

The "perception-only harm" gap is most critical for the threat we examine. No surveyed jurisdiction prohibits real-time synthesis that is viewed but not stored or distributed.

4.6 The Consent Impossibility

Ambient synthesis creates a structural consent impossibility. Traditional consent frameworks assume that consent is theoretically obtainable: a person can be asked, can evaluate the request, and can grant or withhold permission. Ambient synthesis eliminates this possibility through three mechanisms:

First, one cannot opt out of being perceived. To exist in shared physical space is to be visible to others present in that space.

Second, one cannot detect when synthesis is occurring. AR glasses increasingly resemble ordinary eyewear; the Meta Ray-Ban form factor is visually indistinguishable from conventional glasses at conversational distance [5][6]. No reliable indicator reveals whether an observer is running synthesis models.

Third, one cannot prevent technology possession. AR glasses are legal consumer products. Nudification models are freely distributed. No license, registration, or gatekeeping prevents acquisition of the necessary components.

Addressing Technical Opt-Out Proposals

Reviewers may note that technical countermeasures are conceivable. Infrared beacons worn by potential targets could blind cameras. Adversarial patterns on clothing could disrupt person detection models. "Do not process" watermarks could signal opt-out preferences to compliant systems.

We acknowledge these possibilities and reject them on justice grounds, not technical grounds.

Such countermeasures shift the burden of harm prevention entirely to potential victims. Women would be required to wear defensive technology, purchase adversarial clothing, or otherwise armor themselves against unwanted synthesis merely to exist in public space. This burden-shifting is structurally analogous to arguing that assault is preventable if potential victims never venture outside alone: technically accurate, morally bankrupt. The disanalogy with protective equipment requirements (helmets, seatbelts) is categorical: such measures guard against impersonal physical risk, not deliberate misconduct by other humans. Requiring defensive technology against attackers is analogous to mandating tracking devices for women to prevent assault rather than prohibiting assault itself.

The question is not whether defensive measures are possible but whether requiring them is just. A framework that demands women purchase and wear anti-synthesis technology to participate in public life has already conceded that public space is hostile territory requiring defensive preparation. This concession is itself a harm, regardless of whether any individual synthesis occurs.

Moreover, such countermeasures address only compliant systems. Adversarial actors can disable opt-out recognition, modify models to ignore watermarks, or use cameras immune to IR jamming. The asymmetry between universal defensive burden and trivial offensive circumvention renders technical opt-outs inadequate as policy response.

We therefore maintain that consent is structurally unobtainable without imposing unjust defensive burdens on all persons existing in public space. Policy responses must prohibit the conduct itself rather than requiring universal victim armoring.

4.7 Population-Level Harm: The Existence Threat

Prior sections examined harm to individuals who are targeted by ambient synthesis. This section documents a distinct harm category: population-level psychological harm caused by the technology's existence, independent of individual victimization.

4.7.1 Chilling Effects as Established Harm Mechanism

Research on technology-facilitated gender-based violence (TFGBV) documents that fear of potential targeting produces behavioral modification across populations, not merely among those directly victimized.

The International Center for Journalists' global study found that 30% of women journalists self-censor on social media, 20% withdraw entirely, and 18% avoid audience engagement due to fear of online violence [71]. A study of 3,000 Swedish journalists found that 37.3% refrained from reporting on certain topics and 48.1% adapted their reporting to avoid harassment; critically, the authors found that "health-related consequences, both experienced and anticipated, emerge as salient drivers... journalists appear to anticipate emotional costs and adjust preemptively" [72]. Research on UK adults found women significantly less comfortable expressing opinions online due to heightened safety fears [75].

These chilling effects operate through anticipatory threat perception, not actual victimization. The technology's existence, combined with knowledge that targeting is possible, produces behavioral change.

4.7.2 The Grok Incident as Empirical Validation

The January 2026 Grok incident provides direct evidence of ambient threat harm from AI nudification. Within days of users discovering Grok would generate non-consensual intimate imagery on request, coordinated communities used the tool to target hundreds of thousands of women [23][24][25].

Documentation of the aftermath reveals population-level response:

Reuters reported 102 "undressing" requests in a ten-minute span, most targeting unknown women rather than public figures [24]. The Washington Post documented women becoming "afraid to post photos" [76]. The Conversation reported that "the public distribution of these images exerts control over how women choose to present themselves online... someone felt entitled to sexualize your image, instructing Grok to remove clothing and reduce you to a mere body without your consent" [77].

These responses occurred among women who were not individually targeted. The harm was not "my image was synthesized" but "images like mine could be synthesized at any time." This is anticipatory threat, not reactive trauma.

4.7.3 Collective and Cultural Harm

Legal scholars distinguish individual from collective harms in the NCII context. Martin, analyzing Australia's eSafety framework, argues that "beyond individual harms, IBSA and deepfake abuse poses significant 'cultural' and collective harms to women 'as a group'... this issue deprives women of their liberty, capacity to self-determine, and ability to exercise sexual agency and autonomy over their lives, bodies, likenesses, and faces" [78].

This collective harm affects "all members of the same group (here typically women) who live in that society," producing what legal scholar Mary Anne Franks terms deprivation of equal membership: "when cyber harassment persists unchecked, women are not treated as equal members of society" [78].

Chapman and Williams document how the proliferation of undressing applications normalizes the practice, finding that such tools "facilitate and encourage creation of NCII, normalize women's objectification, contribute to a culture in which women's privacy and autonomy are undermined" [79]. Seventy-four percent of deepfake pornography consumers report feeling no guilt [78]. As normalization proceeds, social and moral barriers to abuse erode, making all women potential targets of an increasingly mundane practice.

4.7.4 The Panopticon Mechanism

Feminist surveillance theory provides a framework for understanding how technology existence disciplines behavior through internalized threat perception.

Singh applies Foucault's panopticon to gendered technology contexts: "The possibility of judgment, social, professional, or legal, is enough to ensure compliance... Women learn to shrink their bodies and desire to fit patriarchal ideals through internalized fear of social judgment... This is control not through force, but through perception: the invisible architecture of the digital Panopticon" [80].

Women do not need to be deepfaked to be disciplined by the technology's existence. Knowing that deepfakes could exist is sufficient. The technology creates what Bartky terms a "panoptic male connoisseur" residing in women's consciousness, producing self-surveillance and behavioral modification without any individual act of targeting [80].

4.7.5 Rational Fear vs. Paranoid Anxiety

A critical distinction: women's fear of AI nudification is rational threat assessment, not irrational anxiety. The Grok incident demonstrates that the technology is real and accessible, targeting is random and requires no technical skill, women have documented experiencing this harm, and platforms have not prevented it.

When accurate threat modeling produces the same phenomenological experience as paranoid ideation, the technology has corrupted the epistemic environment. This phenomenon has been observed in other surveillance contexts. Research on mass surveillance documents that "the knowledge, or even the mere perception, that one might be watched influences decisions... The psychological weight of constant observation contributes not only to compliance but to emotional fatigue and diminished psychological resilience" [81].

The parallel to clinical paranoia is instructive. Paranoid delusions are characterized by unfounded beliefs about surveillance or persecution. When surveillance is real, the beliefs are not unfounded; they are accurate. The technology makes the paranoid worldview correct, eliminating the distinction between pathological anxiety and rational fear.

4.7.6 Stage 4 Escalation

Ambient AR synthesis escalates documented Stage 3 harms through three mechanisms:

Spatial extension: Stage 3 harms operate in digital spaces (social media platforms). Stage 4 extends the threat to physical public space. The chilling effect shifts from "fear of posting photographs" to "fear of existing where observers are present."

Temporal acceleration: Stage 3 involves asynchronous harm (photograph uploaded, later discovered and synthesized). Stage 4 involves real-time harm (synthesis occurs as one is observed). The threat is continuous rather than episodic.

Perceptual violation: Stage 3 alters recorded images. Stage 4 alters shared perceptual reality itself. The violation occurs not in an archive but in the lived experience of co-presence.

These escalations do not create new harm categories but intensify documented harms. The critical escalation in Stage 4 is the impossibility of detection. Unlike Stage 3, where a victim might discover a synthesized image, Stage 4 victims can never know if they are being violated in the moment. This epistemic uncertainty forces a permanent, pre-emptive modification of public behavior, panoptic discipline imposed not by the certainty of surveillance, but by the inability to ever verify its absence.

---

5. Mitigation Analysis

5.1 On-Device Safety Filters

The most commonly proposed mitigation is on-device safety filtering. Google's SafetyCore, introduced in November 2024, scans images for nudity and blurs sensitive content locally [64].

However, SafetyCore has been demonstrated to be vulnerable to adversarial attack [14][15]. The Skyld AI team documented a complete attack chain:

1. Model extraction from Android devices via standard static analysis
2. Conversion to PyTorch using publicly available tools
3. Projected Gradient Descent adversarial attacks generating imperceptibly perturbed inputs

The attack requires fewer than 30 lines of Python code and executes in under one minute [14][15].

5.2 Fundamental Limitations of Neural Network Filters

The SafetyCore vulnerability reflects fundamental limitations of deep neural networks established in the adversarial ML literature.

Goodfellow et al. [27] demonstrated that neural networks are susceptible to adversarial examples: inputs imperceptibly different from normal inputs that produce arbitrarily different outputs. Carlini and Wagner [28] showed that adversarial examples are robust across attack scenarios. Papernot et al. [29] demonstrated transferability across models.

For on-device filters, the vulnerability is compounded by the deployment model. Any filter on consumer hardware can be extracted through reverse engineering. Once extracted, it can be attacked. This asymmetry fundamentally favors attackers.

5.3 Defense-in-Depth Considerations

Reviewers may ask whether layered defenses could succeed where individual defenses fail. We address this directly.

Defense-in-depth assumes that attackers must bypass multiple independent defenses. However, the defenses available for this threat are not independent:

- **Platform restrictions** can be bypassed by using open-source models locally
- **On-device filters** can be bypassed via adversarial attacks once models are extracted
- **Hardware restrictions** can be bypassed on rooted/modified devices

An attacker who defeats one defense (moving to local processing) automatically bypasses others (platform moderation). The defenses are serially dependent, not independent layers.

Moreover, the fundamental asymmetry remains: defenders must succeed perfectly against all attack variants; attackers need succeed only once. This asymmetry is well-established in security research and applies fully here.

Critically, the danger is not merely the difficulty of bypass but its distributability. Once one skilled attacker defeats a safety filter, they can package the exploit as a "jailbreak script" requiring zero technical skill to deploy. The asymmetry is not one attacker versus one defender; it is one skilled attacker enabling millions of unskilled users. Friction-based defenses that deter casual abuse collapse entirely once bypass tools circulate.

5.4 Platform and Hardware Restrictions

Platform-level restrictions face the fundamental limitation that once models are released, they cannot be recalled. Gibson et al. document that nudification platforms thrive despite terms-of-service restrictions [1]. Reporting indicates Hugging Face hosts thousands of non-consensual models despite policies [37].

Hardware restrictions (LED indicators, secure enclaves) can be defeated. Investigative reporting documents LED defeat techniques [5][65]. The SafetyCore case demonstrates that system-level services can have models extracted [14][15].

5.5 Mitigation Summary

No technical mitigation we have examined is sufficient to prevent ambient non-consensual nudity rendering by motivated attackers. Mitigations may reduce casual abuse but cannot prevent determined attackers. Policy responses independent of technical prevention are necessary.

---

6. Regulatory and Legal Landscape

6.1 EU AI Act

The EU AI Act (Regulation (EU) 2024/1689), which entered into force August 1, 2024, establishes risk-based AI regulation [30].

Relevant Provisions:

Article 5(1) prohibits certain AI practices. Article 5(1)(c) prohibits:

"AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage"

Article 5(1)(f) prohibits:

"AI systems that infer emotions of a natural person in the areas of workplace and education institutions"

Critical Gaps:

The Act does not prohibit:

- Real-time synthesis of intimate imagery via consumer devices
- On-device AI processing for non-consensual purposes
- Undressing or sexualization models specifically

The biometric provisions (Article 5(1)(c)) address database creation, not real-time perception manipulation. The emotion inference prohibition (Article 5(1)(f)) is context-limited to workplaces and education.

No provision requires disclosure of AI model capabilities on consumer AR hardware.

6.2 US Federal Law

DEFIANCE Act (H.R. 3562)

The Disrupt Explicit Forged Images and Non-Consensual Edits Act passed the Senate in January 2026 [31]. Key provisions:

- Creates federal civil cause of action for victims of non-consensual intimate deepfakes
- Covers "digital forgery" defined as AI-generated imagery depicting identifiable individuals
- Allows statutory damages up to $150,000

Limitation: The Act focuses on distributed content and civil remedies. It does not criminalize creation without distribution, and enforcement requires the victim to become aware and pursue litigation.

TAKE IT DOWN Act (May 2025)

This Act establishes criminal penalties for non-consensual intimate imagery and platform takedown obligations [32].

Limitation: Like the DEFIANCE Act, it assumes distribution as the harm vector. Perception-only synthesis is not addressed.

Section 230 Implications

Section 230 of the Communications Decency Act immunizes platforms from liability for user-generated content. This could immunize AR hardware manufacturers from liability for user-generated synthetic content, though the scope is untested.

6.3 UK Law

Online Safety Act 2023

Section 66B addresses intimate image abuse including AI-generated imagery [33]:

"It is an offence for a person (A) to share an intimate photograph or film of another person (B) without B's consent, where A does so with the intent to cause B distress, and where sharing it results in serious distress to B."

The Act explicitly covers imagery that "appears to show" the victim, capturing deepfakes and AI-generated content.

Limitation: The Act requires "sharing"; perception-only violations without sharing are not covered.

Voyeurism (Offences) Act 2019

As analyzed in Section 4.3, the statutory language targeting equipment "beneath the clothing" does not cover synthesis from clothed imagery [59].

6.4 International Frameworks

GDPR

Articles 4(14) and 9 address biometric data protection [66]. Biometric data is defined as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person."

Synthetic intimate imagery arguably involves processing of physical characteristics, but the GDPR framework focuses on data protection (collection, storage, processing) rather than real-time perceptual manipulation.

No Binding International Framework

No international treaty or framework specifically addresses non-consensual intimate image synthesis. The Council of Europe's Istanbul Convention addresses gender-based violence but predates AI-generated imagery [67].

6.5 Enforcement Challenges

Even with strengthened law, enforcement faces practical barriers:

1. **No server logs**: On-device processing leaves no centralized record
2. **No platform to moderate**: The attack requires no third-party service
3. **Proof difficulties**: Demonstrating synthesis occurred in another's visual field requires evidence that may not exist
4. **Jurisdictional complexity**: Hardware, models, and operation may span multiple jurisdictions

---

7. Discussion

7.1 Why This Threat Was Not Anticipated

The synthesis gap results primarily from research community silos: AR privacy, synthetic media, nudification, and edge AI researchers operate in separate communities with distinct publication venues [1][21][50]. Secondary factors include the privacy-preserving framing of edge AI, which directed attention away from local processing enabling undetectable harms, and commercial incentives favoring beneficial use case narratives over abuse potential analysis.

7.2 Historical Precedent

The pattern of harmful technology applications preceding regulatory response is well-established. Camera phones enabled upskirting before targeted laws emerged [65][68]; similar gaps preceded regulatory responses to social media harassment and political deepfakes.[1] In each case, harm normalized during the gap between capability and response. The challenge is to compress this gap for ambient AR synthesis.

[1] See [20] on social media harassment normalization; [19][56] on deepfake regulatory lag.

7.3 The Timeline Problem

Technical analysis suggests cloud-assisted ambient synthesis is feasible today; edge-only synthesis will be feasible within 12-24 months.

Legislative processes typically require 3-5 years. Regulatory rulemaking requires 1-2 years.

This structural mismatch means the threat will mature before regulatory frameworks respond. Industry self-regulation, platform policies, professional norms, and civil society pressure must fill the gap.

---

8. Recommendations

8.1 Technical (Immediate)

1. **Cross-domain threat modeling workshops**: Convene SNEACI researchers [1][20], XR privacy scholars [50][21], and video generation experts [2][69] for joint threat analysis.
2. **Adversarial robustness research**: Prioritize robustness for wearable safety systems, building on SafetyCore findings [14][15][27][28].
3. **Model extraction detection**: Research extraction detection and prevention for wearable platforms.

8.2 Policy (Near-term)

1. **Extend SNEACI statutes**: Explicitly criminalize non-consensual real-time synthesis, not merely distribution. Model language:

"It shall be unlawful to render, generate, or synthesize intimate imagery of an identifiable person without that person's consent, regardless of whether such imagery is stored, distributed, or shared."

2. **Mandatory capability disclosure**: Require AR manufacturers to disclose what AI model classes can run locally and what safety features exist.
3. **Develop "perceptual consent" doctrine**: Legal scholarship developing this concept for potential legislative adoption.

8.3 Research Agenda (Medium-term)

1. **Formal theory of perceptual consent**: Ground in VR/AR privacy [9][57] and SNEACI harm research [55][20].
2. **Detection mechanisms**: Investigate hardware and OS-level detection for synthetic rendering [19][70].
3. **Prevalence research**: Empirical studies on psychological impacts of perceptual consent violations.

8.4 Industry Standards (Long-term)

1. **Baseline security requirements**: Define security and safety requirements for AR glasses [50][70].
2. **Platform coordination**: Coordinate model-hosting platforms, AR vendors, and regulators on distribution restrictions for wearable-optimized nudification models [1][37][38].

3. **Norm development**: Industry development of perceptual consent indicators or protocols, acknowledging technical limitations.

---

9. Conclusion

This paper has identified and analyzed ambient non-consensual intimate image synthesis via AR wearables. We have shown that component technologies are mature or rapidly maturing, and that their convergence is feasible today in cloud-assisted configurations and will be feasible on-device within 12-24 months across all analyzed scenarios.

We have introduced perceptual consent as a concept naming what is violated when someone synthesizes intimate imagery in their visual field without the subject's knowledge. We have grounded this concept in established frameworks of bodily autonomy [16], informational self-determination [17], and human dignity [18], and demonstrated that it identifies a novel harm category outside existing legal frameworks. We have established principled criteria distinguishing rights-violating synthesis from mere imagination based on realism, externalization, and specificity.

We have documented that population-level psychological harm from technology existence is already empirically established in the Stage 3 (social media deepfake) context, operating through chilling effects, anticipatory threat, collective injury, and panoptic discipline [71][72][77][78][79][80]. Ambient AR synthesis escalates these documented harms from digital to physical space, from asynchronous to real-time, and from image-based to existence-based vulnerability.

We have shown that proposed technical mitigations are fundamentally insufficient due to adversarial vulnerabilities inherent to neural networks [27][28][29], as demonstrated in attacks on deployed systems [14][15].

We have identified gaps in current regulatory frameworks (EU AI Act [30], US federal law [31][32], UK law [33][59]) and provided recommendations for technical, policy, and research responses.

The threat is not speculative. The Grok incident demonstrated immediate, scaled demand when tools become accessible [23][24][25]. The question is not whether ambient AR-based nudification will be attempted, but when, and whether frameworks will be in place to respond.

We call for immediate cross-domain coordination. The alternative is reactive response after harm has accumulated and normalized, repeating the pattern of prior technology-enabled harms. That pattern need not repeat.

Acknowledgments

The author used Claude Opus 4.5 (Anthropic) as assistive technology to accommodate executive function challenges associated with ADHD and autism. AI assistance was used for organizational tasks, formatting, citation verification, and iterative editing. All intellectual content, threat analysis, framework development, and conclusions are the author's original work.

References

[1] C. Gibson et al., "Analyzing the AI Nudification Application Ecosystem," in Proc. USENIX Security Symposium, 2025.

[2] PixVerse, "PixVerse R1: Next-Generation Real-Time World Model," Technical Report, Jan. 2026. [Online]. Available: https://pixverse.ai/en/blog/pixverse-r1-next-generation-real-time-world-model

[3] "PixVerse R1: Real-Time AI Video Generation Arrives," VP Land, Jan. 2026.

[4] "AIsphere Unveils PixVerse R1, Ushering in Real-Time AI Video Era," TMTPost, Jan. 2026.

[5] "Smart glasses: 'I was secretly filmed and trolled online,'" BBC News, Jan. 2026.

[6] "Meta's AI-Powered Smart Glasses Raises Privacy & User Data Concerns," Carleton University, 2025.

[7] "The Role of Edge AI in Real-Time Analytics in 2026 Explained," Kanerika, 2026.

[8] "How Edge AI Enables Real-Time Video Processing in Smart Cameras," TechNexion, 2025.

[9] "Virtual Reality Data and Its Privacy Regulatory Challenges: A Call to Move Beyond Text-Based Informed Consent," California Law Review, 2023.

[10] "Speculative Privacy Concerns About AR Glasses Data Collection," Proc. Privacy Enhancing Technologies (POPETS), 2023.

[11] "SoK: Data Privacy in Virtual Reality," arXiv:2301.05940, 2023.

[12] ShengShu Technology and Tsinghua University, "TurboDiffusion: Accelerating Video Diffusion Models by 100-200 Times," arXiv:2512.16093, Dec. 2025. [Online]. Available: https://arxiv.org/abs/2512.16093

[13] "SnapGen-V: Generating a Five-Second Video within Five Seconds on a Mobile Device," arXiv:2412.10494, 2024.

[14] "When On-Device AI Becomes a Security Flaw: The SafetyCore Case," Skyld AI, 2025.

[15] "Breaking SafetyCore: Exploring the Risks of On-Device AI Deployment," arXiv:2509.06371, 2025.

[16] M. C. Nussbaum, Women and Human Development: The Capabilities Approach. Cambridge University Press, 2000.

[17] D. J. Solove, "A Taxonomy of Privacy," University of Pennsylvania Law Review, vol. 154, no. 3, pp. 477-564, 2006.

[18] I. Kant, Groundwork of the Metaphysics of Morals, 1785. M. Gregor, Trans., Cambridge University Press, 1998.

[19] "Unmasking digital deceptions: An integrative review of deepfake detection," PMC, 2025.

[20] "'There Are No Limits!': AI Undressing Apps and the Normalization of Non-Consensual Intimate Imagery," PubMed, 2025.

[21] "A systematic literature review on Virtual Reality and Augmented Reality in terms of privacy, authorization and data-leaks," arXiv:2212.04621, 2022.

[22] "Exploring the Uncoordinated Privacy Protections of Eye Tracking and VR Motion Data," arXiv:2411.12766, 2024.

[23] "Behind Grok's mass digital undressing lies an 'unsurprising' cohort," ABC News Australia, Jan. 2026.

[24] "Attorney General Bonta Launches Investigation into xAI, Grok," California Attorney General Press Release, Jan. 2026.

[25] "Grok-created images of real people in bikinis, underwear banned on X," Mashable, Jan. 2026.

[26] A. Shostack, Threat Modeling: Designing for Security. Wiley, 2014.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. ICLR, 2015. arXiv:1412.6572.

[28] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in Proc. IEEE S&P, 2017.

[29] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in Proc. IEEE Euro S&P, 2016.

[30] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," Official Journal of the European Union, Aug. 2024. EUR-Lex: 32024R1689.

[31] "DEFIANCE Act of 2025," H.R. 3562, 119th Congress (2025-2026). [Online]. Available: https://www.congress.gov/bill/119th-congress/house-bill/3562

[32] "TAKE IT DOWN Act," Pub. L. No. 119-XX, May 2025.

[33] "Online Safety Act 2023," UK Public General Acts, c. 50, Section 66B.

[34] N. Henry et al., "Image-Based Sexual Abuse: A Study on the Causes and Consequences of Non-Consensual Nude or Sexual Imagery," 2020.

[35] "Prevalence and Impacts of Image-Based Sexual Abuse," Cyber Civil Rights Initiative, 2025.

[36] "Development of an AI-powered AR glasses system for real-time first aid assistance," PMC, 2025.

[37] E. Maiberg, "Hugging Face Is Hosting 5,000 Nonconsensual AI Models of Real People," 404 Media, July 15, 2025. [Online]. Available: https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/

[38] "UnfilteredAI," Hugging Face Model Repository.

[39] "Researchers warn of rise in AI-created non-consensual explicit images," University of Florida News, May 2025.

[40] "The 10 Best AI Video Generators for Content Creation," Digital Bricks, 2025.

[41] "The Best AI Video Creation Trends from 2025," Clippie AI, 2026.

[42] "PixVerse V5.5: Multi-Shot Video with Sound," SuperMaker AI, 2025.

[43] "PixVerse-R1 LLM Benchmark Data," Automatio AI, 2026.

[44] "Meta AI Glasses: Revolutionary AR Technology or Privacy Nightmare?" GigeNet, 2025.

[45] M. Speicher et al., "AR Privacy Concerns in Social Contexts," in Proc. ACM CHI, 2020.

[46] R. Williams et al., "Bystander Privacy in AR," IEEE VR, 2021.

[47] "Edge Computing and AI: The Future of Real-Time Data Processing," Sapien.io, 2025.

[48] "AI-based Wearable Vision Assistance System for the Visually Impaired," arXiv:2412.20059, 2024.

[49] "LoXR: Performance Evaluation of Locally Executing LLMs on XR Devices," arXiv:2502.15761, 2025.

[50] "Immersed in Reality Secured by Design: A Comprehensive Analysis of Security Measures in AR/VR Environments," arXiv:2404.16839, 2024.

[51] "Deepfakes and Synthetic Media," TechUK, 2025.

[52] "The rise and risks of synthetic media," Digital Watch Observatory, 2025.

[53] "2025 AI Safety Index," Future of Life Institute, 2025.

[54] "SDXS: Real-Time One-Step Latent Diffusion Models with Image Conditions," arXiv:2403.16627, 2024.

[55] "'Violation of my body:' Perceptions of AI-generated non-consensual (intimate) imagery," arXiv:2406.05520, 2024.

[56] "Regulating AI Deepfakes and Synthetic Media in the Political Arena," Brennan Center for Justice, 2025.

[57] J. Bailenson and J. Cummings, "VR Consent and Embodiment," Stanford Virtual Human Interaction Lab, 2016.

[58] "Designing Effective Consent Mechanisms for Spontaneous AR Interactions," in Proc. ACM CHI, 2025.

[59] "Voyeurism (Offences) Act 2019," UK Public General Acts, c. 2, Section 67A.

[60] California Penal Code § 647(j)(4).

[61] California Civil Code § 1708.85.

[62] Virginia Code § 18.2-386.2.

[63] Texas Penal Code § 21.165.

[64] "Google forcing Android System SafetyCore on users to scan for nudes," Kaspersky, 2025.

[65] "Smart Glasses: Cool Tech or a Privacy Threat?" Panda Security, 2025.

[66] "General Data Protection Regulation," Regulation (EU) 2016/679, Articles 4(14), 9.

[67] Council of Europe, "Convention on preventing and combating violence against women and domestic violence (Istanbul Convention)," CETS No. 210, 2011.

[68] "Not a Good Look, AI: What Happens to Privacy When Glasses Get Smart," Cyber Security Advisors Network, May 2025.

[69] "Owl-1: Omni World Model for Consistent Long Video Generation," arXiv:2412.09600, 2024.

[70] "Augmenting Security and Privacy in the Virtual Realm: An Analysis of Extended Reality Devices," arXiv:2402.03114, 2024.

[71] International Center for Journalists, "The Chilling: A Global Study of Online Violence Against Women Journalists," 2023. [Online]. Available: https://www.icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists

[72] M. Björkenfeldt et al., "Self-Censorship Among Journalists: Anticipatory Threat and Professional Adaptation," Digital Journalism, 2025. DOI: 10.1080/21670811.2025.2601961

[73] WITNESS, "Technology-Facilitated Gender-Based Violence," Mar. 2025. [Online]. Available: https://blog.witness.org/2025/03/technology-facilitated-gender-based-violence/

[74] UNFPA, "An Infographic Guide to Technology-Facilitated Gender-Based Violence," 2025. [Online]. Available: https://www.unfpa.org/

[75] L. Lumsden et al., "Gender and Online Expression: A UK Population Study," arXiv:2403.19037, 2024.

[76] Washington Post, "X Users Tell Grok to 'Undress' Women in Their Photos," Jan. 6, 2026.

[77] The Conversation, "How AI-Generated Sexual Images Cause Real Harm, Even Though We Know They Are Fake," Jan. 15, 2026. [Online]. Available: https://theconversation.com/how-ai-generated-sexual-images-cause-real-harm-even-though-we-know-they-are-fake-273427

[78] F. Martin, "Online Safety Regulation of Deepfake and Image-Based Sexual Abuse in Australia," Alternative Law Journal, 2025. DOI: 10.1080/10383441.2025.2504791

[79] A. Chapman and M. Williams, "'There Are No Limits!': AI Undressing Apps and the Normalization of Nonconsensual Intimate Deepfakes," Violence Against Women, 2025. DOI: 10.1177/10778012251397966

[80] A. Singh, "The Gendered Panopticon: How Surveillance Enforces Binary Norms," Nickled and Dimed, Apr. 2025. [Online]. Available: https://nickledanddimed.com/2025/04/21/the-gendered-panopticon-how-surveillance-enforces-binary-norms/

[81] "The Panopticon 2.0: Mass Surveillance in the Digital Age," The Uncensored Truth, 2025. [Online]. Available: https://theuncensoredtruth.com/part-1-the-panopticon-2-0-mass-surveillance-in-the-digital-age/

[82] M. Vef and B. Groß, "NUCA: AI Camera That Redefines Image Creation," 2024. [Online]. Available: https://nuca.rocks/

Corresponding author: Travis Gilly, Real Safety AI Foundation, t.gilly@ai-literacy-labs.org