Position: AI/ML Deepfake Research is Misaligned with AI-Generated Non-Consensual Intimate Imagery (AIG-NCII)

Li Qiwei [1] **Wells Lucas Santo** [1] **Sarita Schoenebeck** [1] **Eric Gilbert** [1]

Abstract

AI-generated non-consensual intimate imagery (AIG-NCII) is not adequately addressed in AI/ML literature regarding AI-generated media, commonly referred to as "deepfakes". While research on deepfakes currently focuses on its epistemic harms—or harms relating to truth and authenticity—this is misaligned with the dominant reality of generative AI abuse involving sexualized imagery. We conduct a landscape analysis of highly-cited works to demonstrate that technical interventions addressing deepfakes almost entirely ignore AIG-NCII, limiting the research ecosystem to authenticity detection tools. In this position paper, we argue that existing interventions address *viewer-centric* epistemic harms, such as fraud or scams, but ignore *subject-centric* dignity harms, such as AIG-NCII. We illustrate that knowing an image is synthetic does not mitigate harms to subjects and may, in some cases, even exacerbate them. We conclude by offering recommendations to realign the field, including updating threat models to consider subject-centric harms and addressing AIG-NCII in AI safety research. Finally, we caution that researchers should only engage in this high-risk domain if they implement safety guardrails for both subjects and researchers and establish partnerships with domain experts in sexual violence prevention.

Content warning: This paper includes content about online sexual violence and suicide.

[1]School of Information, University of Michigan, Ann Arbor, USA. Correspondence to: Li Qiwei <rrll@umich.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

The creation of non-consensual intimate imagery (NCII) is a concerningly dominant use case of generative AI, whose harms are still inadequately addressed. In a report released in January 2026, AI Forensics found that the majority use case of the commercial AI system "Grok" is undressing people without their consent (Bouchaud, 2026). New York Times and the Center for Countering Digital Hate separately found that after Elon Musk shared an image of an undressed woman superimposed on a SpaceX rocket produced by Grok in late December, Grok was asked to generate 4.4 million images the following week, compared to roughly 311,000 the week before Musk's post (Conger et al., 2026). It is estimated that more than 3 million of these images were sexualized, with at least 23,000 being of children (Center for Countering Digital Hate, 2026). Yet, despite the empirical reality that sexual abuse is a primary driver of generative AI usage, and the growing development of applications with the ability to produce AI-generated non-consensual intimate imagery (AIG-NCII), the phenomenon is largely absent from extant research involving generative AI imagery. Instead, the technical interventions developed by the AI/ML community are largely designed for a different set of needs focused on threats to truth and trust.

While current interventions are designed to address the question of whether a piece of media is authentic or synthetic, the inattention to AIG-NCII has resulted in oversights in how safety tools are developed. Drawing on the fundamental human right of dignity, as outlined in the Universal Declaration of Human Rights from the United Nations (1948), we argue that interventions should also focus on addressing dignity harms that cause injury to the subject.

This position paper argues that there is a structural misalignment between AI/ML research agendas and the reality of AI-generated media, with existing concerns focusing on viewer-centric epistemic harms, which ignore or even exacerbate subject-centric dignity harms. Our contributions in this position paper are as follows:

1. We surface a disconnect between research motivation and the actual harms of deepfakes. Through a systematic landscape analysis of highly-cited works in

top-tier venues between 2020 and 2025, we find scant consideration for cases of AIG-NCII, despite it likely accounting for more than half of generative AI usage (Bouchaud, 2026; Center for Countering Digital Hate, 2026; Security Hero, 2023).

2. We analyze how this lack of engagement has resulted in the proliferation of "authenticity" tools. We demonstrate how these tools are insufficient for AIG-NCII and, in specific deployment contexts, could exacerbate subject-centric dignity harms.

3. Finally, we offer a series of recommendations for researchers and practitioners to realign technical interventions to account for AIG-NCII. We stress that all researchers who work in this domain must partner with domain experts in online sexual violence and utilize threat models that consider subject-centric dignity harms in technical interventions.

2. The absence of AIG-NCII in harm reduction research

A "deepfake" is a colloquial term used to describe AI-generated or altered content (Diel et al., 2025). This includes deceptive political videos such as that of using the likeness of President Biden to tell voters not to vote in the New Hampshire primary during the 2024 election (Bond, 2024), or fraud content such as using the likeness of Elon Musk to offer an investment opportunity that led to billions of dollars in monetary losses (New et al., 2024). A body of technical research has been developed to address deepfakes. As we show in our analysis, the vast majority of this work has been conducted without acknowledging harms of a sexual nature. While not all deepfake content is categorized as AIG-NCII, it still exists as a dominant form of deepfake, with reports estimating that up to 98% of deepfake videos are pornographic in nature (Security Hero, 2023). In fact, the term "deepfake" originates from a direct reference to sexual harm, with the term being derived from the username of a Reddit user who shared custom-made AI-generated videos depicting actresses performing sexual acts (Burkell & Gosse, 2019). Since then, however, the word has entered the global lexicon to refer to a broad range of AI-generated content, including AI-generated misinformation, scams, and political images and videos.

2.1. AIG-NCII as a distinct harm category

We use the acronym AIG-NCII (AI-generated non-consensual intimate imagery) to refer to the phenomenon of sexualized deepfake content of a specific individual that exists without their consent. This can refer to synthetic content that is created with generative AI technology to "nudify" or "undress" a subject without their explicit agreement (Van der Nagel, 2020)[1]. Other terms that describe this same phenomenon include "AI-NCII", "AI-generated image-based sexual abuse" (Henry et al., 2026), synthetic nonconsensual explicit imagery (SNCEI) (Wei et al., 2025), and "deepfake pornography" (Furizal et al., 2025). AIG-NCII as a concept is derived from traditional non-consensual intimate imagery (NCII)[2]. AIG-NCII is highly gendered, with the vast majority of those impacted being women and girls (Security Hero, 2023; Bouchaud, 2026). At a societal level, AIG-NCII represents attempts to silence, de-platform, and de-legitimize agency both online and offline (Maddocks, 2020).

From GANs to diffusion models. The technical barriers to creating AIG-NCII have lowered precipitously over the last decade. Initially, from approximately 2017 to 2022, face-swapping was the primary mechanism used to create AIG-NCII, by superimposing a face onto another person's body. This was enabled by autoencoder architectures, popularized by open-source repositories like DeepFaceLab (Perov et al., 2020). The infamous Deepnude application, which "undressed" women in images, relied directly on the Pix2Pix conditional GAN architecture (Isola et al., 2017; Wang et al., 2018). The current phase of AIG-NCII creation is driven by diffusion-based synthesis, enabled by the release of open-weight latent diffusion models such as Stable Diffusion (Rombach et al., 2022; Schuhmann et al., 2022). Unlike face-swapping, diffusion models allow for the generation of sexualized imagery via text-to-image prompting. Techniques such as Dreambooth (Ruiz et al., 2023) and Low-Rank Adaptation (LoRA) (Hu et al., 2022) are now standard tools in AIG-NCII communities to fine-tune a specific individual's likeness with few reference photos.

Academic research directly contributes to AIG-NCII. Han et al. (2025) found in an analysis of the online community for Mr. DeepFakes, a prominent marketplace for deepfake content, that there was a "significant sharing of academic work" on its website, with direct references to GitHub repositories for deepfake tools that cite 43 academic papers. Many of these deepfake tools are forked from open-source models directly from research papers, and many of these applications are simply wrappers around open-source research code (Han et al., 2025; Gibson et al., 2025). Additionally, the use of nude bodies as training data, often collected without consent, gives rise to these capabilities (Cintaqia et al., 2025).

Limitations of law and policy. Despite legal prohibitions in the U.S. and abroad, enforcement remains insufficient,

[1]It is important to note here that AIG-NCII does not require the subject to be fully nude (Batool et al., 2024). Recent examples of AIG-NCII have included content that has attempted to remove hijabs from women (Tenbarge, 2026).

[2]Non-consensual intimate imagery (NCII) is also colloquially known as "revenge pornography".

costly, and reactive, often placing the burden of discovery on the victim (Sen. Durbin, 2024; Qiwei et al., 2025; Congress.gov, 2025). For example, while the U.K. Online Safety Act criminalizes sharing AIG-NCII, abuse has merely shifted to non-compliant platforms and encrypted apps like Telegram (tel, 2024). Similarly, South Korea responded to major scandals (Gibson, 2019; Bicker, 2020) by criminalizing possession (Yim, 2024), yet deepfake generation tools remain accessible. Moderation efforts on individual platforms face a similar "whac-a-mole" dynamic (Ding et al., 2026). When CivitAI banned real-person likeness models (CivitAI, 2025; Maiberg, 2025a; Wagner & Cetinic, 2025), the models simply migrated to HuggingFace (Maiberg, 2025b).

2.2. Landscape analysis of existing literature

To quantify the misalignment between existing research concerns and the reality of AIG-NCII, we conducted an analysis of technical defense papers published between 2020 and 2025. Our analysis reveals that the literature addressing deepfake harms almost entirely ignores AIG-NCII.

Methodology. Our goal was to locate works that aimed to address harms regarding deepfakes or synthetic media more broadly. We queried Google Scholar for papers containing a specific set of keywords, using the following query: *("detection" OR "detector" OR "forensics" OR "recognition" OR "watermark") AND ("deepfake" OR "synthetic image" OR "fake image" OR "diffusion")*. This initial search criteria yielded 965 papers. Next, we filtered our results to papers published only at the top-tier venues: *"CVPR" OR "ICCV" OR "ECCV" OR "NeurIPS" OR "ICML" OR "ICLR"*. We excluded workshop papers and included arXiv preprints that were returned in our search. Papers with more than 80 citations from other venues were also included. This brought our resulting dataset to 379 papers. Finally, we filtered the results down to the top 100 most cited papers, and manually excluded papers where diffusion models were utilized for unrelated computer vision tasks such as detecting tumors, detecting cars, or detecting cracks in steel. We also excluded one retracted paper. This process left a final dataset of 39 papers that we qualitatively analyzed. We examined each paper for engagement with AIG-NCII, with particular attention to the usage of the following terms: *"non-consensual intimate imagery" "NCII", "revenge porn","sexual violence" "porn", "nudity", "undress","obscene"*. See Table 2 in the Appendix for the final list of 39 papers.

Results. We categorized the 39 papers into three tiers of engagement with AIG-NCII.

1. **No mention (34 papers):** The paper frames the problem exclusively as misinformation, fraud, or technical artifact detection.
2. **Mention only (5 papers):** The authors reference AIG-NCII terms in passing, such as in the introduction or broad impact statement, but the proposed technical method remains generic.
3. **Technical implementation (0 papers):** The authors design an intervention with a threat model specific to AIG-NCII.

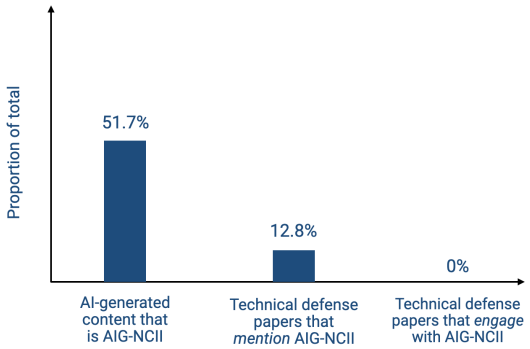

Figure 1. Proportion of AI-generated content that is classified as AIG-NCII (Bouchaud, 2026) compared to the proportion of technical defense papers that merely mention AIG-NCII. No papers found in our landscape analysis meaningfully engaged with AIG-NCII.

Our analysis affirms existing research stating that deepfake research is overwhelmingly motivated by harms related to trust, fraud, and political misinformation (Rini, 2020; Harris, 2021).

3. Review of authenticity-based interventions

In order to deal with harms pertaining to truth, our analysis found that the AI/ML community has coalesced around a series of tools designed to distinguish *synthetic* from *authentic* media. We categorize these technical interventions into three primary paradigms: detection, provenance, and watermarking. While they differ significantly in their implementation, they share one foundational assumption, that *truth-verification, or authenticity, is the primary proxy for safety.* In this section, we review these three paradigms and note the assumptions that underpin their design.

3.1. Detection

Current AI detection attempts to approximate a classification problem, wherein a model learns a decision boundary between the distribution of authentic media and synthetic media. The primary assumption of this paradigm is that a robust decision boundary exists, and that successfully determining the boundary between authentic and synthetic is a sufficient condition for addressing harm.

While earlier literature focused on identifying GAN-specific artifacts in the media asset (Frank et al., 2020), the prolifer-

ation of diffusion models has required a fundamental shift in the feature extraction process for identifying synthetic content. More recent detection methods target the unique fingerprints introduced by the iterative de-noising process of diffusion models. For example, DIRE (Wang et al., 2023) utilizes the observation that diffusion-generated synthetic images show lower reconstruction error inverted through a pre-trained diffusion process as compared to authentic images. Similarly, Corvi et al. (2023) identified and analyzed distinct spectral traces left by the Gaussian noise scheduling that are inherent to latent diffusion models. To address the rapid evolution of generator architectures (e.g., Stable Diffusion (Rombach et al., 2022) and FLUX (Black Forest Labs, 2024)), research has increasingly moved utilizing feature spaces from foundational vision-language models like CLIP (Radford et al., 2021) to identify synthetic semantic patterns that generalize across architectures (Ojha et al., 2023).

3.2. Provenance

Provenance methods, as defined by the Coalition for Content Provenance and Authenticity (C2PA) (2026) technical specification, attempt to establish a history of modifications for a media asset. Unlike methods in detection, this paradigm for protection relies on cryptographically verifiable information that can be used to verify that an asset is free from tampering. At each point of asset creation or modification, a digital signature is bound to a hash of the pixel data of the asset, along with a manifest containing metadata assertions, such as content ownership and timestamp (Rosenthol, 2022). The assumption guiding provenance methods is that having a verifiable chain-of-custody resolves trust in where a piece of media originated, and whether it was edited along the way. In other words, interventions following the provenance paradigm are concerned with tracking the lineage from a source asset to any of its modified outputs, such that the lineage itself is an indicator of authenticity.

3.3. Watermarking

Content watermarking techniques aim to embed signals invisible to the human eye directly onto the media content at generation, to signal that content as synthetic. Unlike metadata, which can be easily stripped and removed from the asset, watermarks aim to be more robust against being removed, even with cropping, filters, and other image transformations. Approaches within this paradigm include latent watermarking (Fernandez et al., 2023) and sampling-based watermarking (Wen et al., 2023). Industry implementations, such as Google's SynthID (DeepMind, 2025), embeds signals onto the media that are detectable only when paired with a specialized detector model or decoding mechanism. Watermarking is often used to help with post-hoc detection.

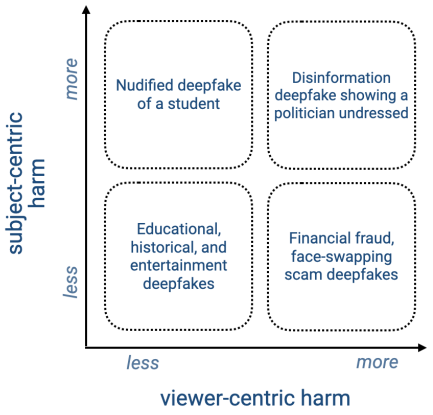

Figure 2. Examples of deepfakes that cause harm to the viewer, subject, or both parties.

4. Consequences of ignoring AIG-NCII

The intervention paradigms around detection, provenance, and watermarking are meant to restore the viewer's ability to discern what is authentic. For AIG-NCII cases, however, the subject suffers from violation of their dignity regardless of authenticity. In this section, we discuss possible consequences overlooking these dignity violations.

4.1. Neglect of subject-centric dignity harms

The exclusion of AIG-NCII in deepfake harm reduction research has resulted in an ecosystem that solely focuses on epistemic harms, or harms relating to truth and authenticity. However, as Harris (2021) notes, these concerns often overlook the severity of harms to subjects. Following Chesney & Citron (2019) and Citron (2018) and drawing from the UN Universal Declaration of Human Rights (United Nations, 1948), we argue that AIG-NCII violates the fundamental human right of dignity for deepfaked subjects, which is distinct from epistemic harms to viewers. We build upon the framework by Olteanu et al. (2025) to distinguish between harms to the viewer and the subject:

Viewer-centric harms encompass both individual deception and broader societal epistemic degradation. On an individual level, this includes fraud, such as scams mimicking a friend's likeness. However, as Rini (2020) argues, this actually presents a societal harm due to the erosion of authenticity and the destabilization of shared reality (Chesney & Citron, 2019). If one was led to believe false information, they may have suffered a viewer-centric epistemic harm.

Subject-centric harms occur when a person's likeness is utilized without consent, regardless of whether the viewer is deceived. Drawing on Citron (2018)'s idea of sexual privacy, the harm arises from the undignified presentation

of the subject and the resulting loss of autonomy. At a high level, a dignity harm occurs when one's likeness is used in ways they did not *consent* to. We highlight two instances of this harm that map onto specific AI/ML techniques. The first is non-consensual identity preservation, in which a model retains a recognizable likeness of a specific person without their consent. Defenses against this mode may aim to prevent identity retention—for example, by disrupting a model's ability to learn or reproduce a specific subject's features (Van Le et al., 2023). The second is non-consensual modification, in which the subject's likeness is intentionally preserved, but the image is altered in ways they did not consent to, such as removing clothing or face-swapping with pornographic videos.

Viewer-centric and subject-centric harms are orthogonal, as shown in Figure 2. While these harms often overlap—for instance, a political deepfake may deceive constituents (viewer-centric) while damaging the politician's reputation (subject-centric)—the mechanisms required to address them differ. The lack of engagement with subject-centric harms has led to a lapse in defense design, where merely labeling content as "synthetic" does not mitigate subject-centric harms. The misalignment between technical capabilities and safety needs is so acute that even platform governance bodies have begun questioning use of these tools. The Oversight Board for Meta recently noted that "labeling manipulated content is not appropriate in this instance because the harms stem from the sharing and viewing of these images—and not solely from misleading people about their authenticity" (Oversight Board, 2024). Indeed, authenticity-based interventions do not mitigate harms to subjects, especially given that prominent online forums that host AIG-NCII already routinely label the content as fake (Han et al., 2025).

4.2. Authentic $\neq$ safe

The current trajectory of deepfake defense research optimizes for authenticity metrics. When applied to AIG-NCII, this creates a fundamental category error, treating authenticity as a proxy for safety. This substitution fails because, unlike political misinformation where falsehood is the primary harm, sexual violence is defined by the absence of consent, which is violated regardless of whether an image is authentic or synthetic. As illustrated in Table 1, the axis of artificiality (what existing interventions measure) is orthogonal to the axis of consent (what determines safety). Relying on authenticity-based tools creates a blunt instrument that cannot address the case of AIG-NCII, because it conflates non-consensual synthetic imagery with that of consensual synthetic imagery. At the same time, the ecosystem risks building over-censoring tools that stifle legitimate expression while failing to address traditional non-consensual imagery simply because it is "authentic". We argue that until technical interventions can account for this orthogonality, evidence of authenticity should not be treated as sufficient evidence of safety.

	Safe	**Harmful**
Synthetic	Self-expression of artistic nudity using AI	AIG-NCII
Authentic	Consensual pornography	Traditional NCII

Table 1. The orthogonality of harm. The current detection paradigm can only distinguish between synthetic and authentic media. However, safety falls along an orthogonal axis that is determined by consent.

4.3. Potential misuse of authenticity tools

The implicit assumption guiding current research agendas is that safety can be achieved once the most accurate authenticity model is developed. We challenge this assumption. In the following contexts, we demonstrate how authenticity interventions, when built without consideration for a subject-centric threat model, may actually exacerbate harms to subject dignity.

When used by online platforms. Current regulations, such as the EU AI Act (ArtificialIntelligenceAct.eu, 2025), as well as platform policies on Meta (Meta Platforms, 2024) and TikTok (TikTok, 2026), prioritize the labeling of synthetic media that is identified. While effective for viewer-centric epistemic harms, this model of public labeling could backfire for AIG-NCII. A label stating that an image is made with AI does not address the fact that the image is shared on online platforms. While some platforms explicitly ban nudity, AIG-NCII also includes cases where the subject is not fully nude, which is ignored by these regulations (Batool et al., 2024). When labeling content is prioritized over its removal, this could create a perverse outcome where the abuser may be protected from moderation consequences so long as they are transparent about the synthetic, AI-generated nature of the image.

When used by abusers. Research in online sexual violence indicates that perpetrators are often driven by an assertion of power over a victim, rather than by sexual gratification on its own (Henry & Beard, 2024; Henry & Powell, 2016). In fact, Marini et al. (2024) has shown that people are less aroused when they find that an image is identified as AI-generated. As noted by Massanari (2017), these communities are not passive consumers but active participants who "demonstrate technological prowess" in aggregating disparate pieces of content to target and verify the identity of victims. These same communities may use authenticity-based tools to locate content in order to further harass and

dox victim-survivors. In this context, authenticity verification may become a tool that allows users to sort through authentic and synthetic imagery for abuse.

When used against NCII victim-survivors. Finally, the utility of authenticity labels is not consistent for victim-survivors, fluctuating depending on the harm that the subject is dealing with. For victim-survivors of traditional NCII, plausible deniability may actually offer a safety mechanism to protect their dignity. In this context, ambiguity offers a form of protective cover for victim-survivors. If a detection system definitively labels traditional NCII as "authentic", it inadvertently acts as a verification tool for abusers, confirming the victim's exposure to the public. By removing this uncertainty, technical interventions may out victims who could otherwise have maintained some degree of social safety by casting doubt on the image's veracity.

5. Recommendations

Addressing AIG-NCII is an extremely difficult task, and it is one that faces significant barriers both ethically and technically. Given the sensitive nature of this abuse, we cannot always propose definitive solutions. In this section, we raise recommendations to best address the problem. Ultimately, we believe there is a need for both technical and social pathways for addressing the proliferation of AIG-NCII, and this requires coordination between AI/ML researchers and practitioners, social scientists, policy makers, sexual violence prevention experts, and victim-survivor advocates.

R1. Decouple epistemic and dignity harms. As we have shown, care must be taken when applying authenticity markers. If a system identifies potential AIG-NCII, it should not apply a public label, as this keeps content visible, and may exacerbate harm. Instead, detection should serve as a backend flag that triggers precautionary handling (suppression or triage) treating AIG-NCII like traditional NCII when content depicts real people in sexualized contexts. We urge the research community to study the specific trade-offs of labeling AIG-NCII before deploying detection tools. Future work should identify how to verify content in a way that protects the privacy of victims while empowering deepfake subjects to defend themselves. Furthermore, we must weigh the asymmetrical harm of errors in labeling. While temporarily restricting consensual content is often reversible, failing to intervene in situations of sexual violence imposes irreversible harms. Future work must rigorously evaluate how transparency standards designed for misinformation may inadvertently endanger privacy.

R2. Elevate subject-centric dignity harms. The violation of dignity must be elevated to the same tier as political misinformation. Researchers should mirror the shift in computer security towards analyzing intimate partner violence (IPV), where the adversary is personally known to the victim-survivor (Chatterjee et al., 2018; Havron et al., 2019; Freed et al., 2018). In AIG-NCII, the adversary is often an actor equipped with limited reference photos and parameter-efficient fine-tuning techniques. Under this framework, the goal shifts from maximizing detection accuracy to minimizing identity preservation. Defenses are successful if they prevent the reproduction of a specific identity or make it more challenging. Furthermore, we must reject the assumption that publicly available data is synonymous with consensual data. Privacy is violated by the migration of information outside its intended context (Nissenbaum, 2004). While modeling general human features may be necessary in some contexts, the non-consensual ingestion of nude or sexualized bodies crosses an ethical red line (Stark, 2018; Scheuerman et al., 2021). Researchers should abandon the unethical curation of datasets from sensitive domains (Cintaqia et al., 2025).

R3. Restrict high-risk research assets. We call for a re-evaluation of open-release norms for architectures explicitly optimized for high-fidelity identity retention and inpainting. While open science is a core value of the field, an emerging consensus in AI Safety literature recognizes that the risks of open-sourcing highly capable, dual-use models often outweigh the benefits (Widder et al., 2022; Solaiman, 2023; Seger et al., 2023). Models and fine-tuning techniques that demonstrate state-of-the-art performance (such as "cloning" the likeness of a person from a few photographs) should be subject to gated or researcher-only access. By placing additional friction on the tools of creation, we can reduce the size of the threat downstream.

R4. Consider proactive prevention. The field should move beyond post-hoc detection and engage seriously with adversarial defense mechanisms against inpainting, one of the primary techniques for nudification. Adversarial immunization introduces human-imperceptible perturbations that disrupt internal representations in generative models, preventing style mimicry or inpainting (Jeon et al., 2025; Guo et al., 2025; Van Le et al., 2023; Shan et al., 2023). While these defenses can be brittle against transformations (Hönig et al., 2025), recent work may be closing this gap by making perturbations harder to remove (Kim et al., 2026) and dramatically lower the per-image protection cost (Ozden et al., 2025). Even imperfect defenses raise the cost for lay-abusers and disrupt the generative pipelines.

R5. Develop safety-aligned metrics. Success must be measured by harm reduction, not merely detection accuracy. We need metrics that verify whether systems actually reduce the prevalence of abuse. This requires creative evaluation strategies. For instance, Cretu et al. (2025) utilized ethical proxies to assess child-safety filters without generating actual CSAM. This may offer inspiration for evaluating

AIG-NCII interventions without producing harmful content.

R6. Integrate AIG-NCII into AI Safety. Subject-centric harms must be elevated to a core concern within the definition of AI Safety. Scholars have increasingly critiqued the mainstream discourse for focusing on existential risks while excluding present-day harms (Gyevnar & Kasirzadeh, 2025; Ahmed et al., 2023; Hazra et al., 2025). We further note that the dominant model-side safety techniques (e.g., concept erasure and NSFW filtering (Gandikota et al., 2023; Schramowski et al., 2023)) are insufficient for AIG-NCII because they assume a cooperative model operator. The AIG-NCII ecosystem is defined by open models and non-compliant platforms (Gibson et al., 2025). AIG-NCII therefore requires safety research that does not depend on operator goodwill. If the field is to meet its goal of protecting human welfare, the scope of AI Safety must expand to include the immediate violence of AIG-NCII. Conference organizers should recognize the scale of this abuse and allocate the same rigor and resources currently afforded to sub-fields such as algorithmic bias and toxic content generation (Weidinger et al., 2021).

R7. Establish ethical partnerships and guardrails. Research into AIG-NCII poses unique ethical and psychological challenges and is not suitable for all researchers. Guardrails must be implemented to mitigate secondary traumatic stress for those who review sensitive content (Williamson et al., 2020). Crucially, technical researchers must avoid speculatively deriving threat models in isolation. Work must be empirically grounded and co-designed with domain experts in online sexual violence and victim advocacy (Costanza-Chock, 2020). These experts should be integrated as partners during the initial design phase, rather than merely consulted for post-hoc validation.

R8. Account for victim-survivor plurality. Interventions must respect the spectrum of survivor needs. As noted by McGlynn & Westmarland (2019), victim-survivors may prioritize drastically different outcomes, ranging from criminal prosecution to content removal. A one-size-fits-all technical solution cannot serve all these needs. Researchers should leverage frameworks of restorative justice to design flexible tools that respect survivor agency, rather than imposing a monolithic technical "solution" (Schoenebeck et al., 2021).

R9. Acknowledge that social harms require social interventions for remediation. While existing interventions are focused on detecting and identifying the synthetic nature of deepfakes, this does not actually resolve any of the dignity-based harms that have been inflicted on victim-survivors. Experts in interpersonal violence (IPV) and sexual violence reduction should be consulted when developing possible interventions to remediate the harms of AIG-NCII.

6. Alternative Views

In this section, we address the primary objections to our arguments that AI/ML research must account for AIG-NCII.

AV1: Addressing social harms is the domain of law and policy, not AI/ML. One objection to this paper's argument is that AI/ML researchers are only responsible for optimizing technical capabilities, while the regulation of those capabilities belongs to policymakers. From this perspective, AIG-NCII is fundamentally a legal problem that arises from under-resourced legal systems and platforms that fail to enforce abuse. How generative AI tools are used is out of scope for AI/ML researchers.

Response: We agree that law and policy are essential, but we reject the premise that research into technical protections is therefore irrelevant. First, the timing mismatch between AI development and legal enforcement makes relying solely on the law unfeasible. The legal system operates on timescales of years, and generative AI capabilities evolve in weeks. In the example of the UK Online Safety Act, by the time the legislation passed, the dominant mechanism for AIG-NCII had already shifted from face-swapping to diffusion synthesis. The decision to work on defense methods against AIG-NCII is as much a research choice as it is to work on other technical innovations such as LoRA. Much like how the fight against Child Sexual Abuse Material (CSAM) has involved legal, technical, and advocate coordination, we argue that technical protections can work in tandem with legal efforts against AIG-NCII, which requires a multi-pronged approach.

AV2: Prevention of AIG-NCII is technically intractable. Even if researchers accept responsibility, proposed interventions simply do not work. Adversarial defenses such as inpainting perturbation are brittle and easily defeated by compression or model updates (Sun et al., 2023; Guo et al., 2025; Goodfellow et al., 2014; Athalye et al., 2018; Hönig et al., 2025). Furthermore, the sheer scale of the internet makes protecting every user's photo impossible, and sophisticated abusers can easily bypass protections using slightly altered inputs. Therefore, proposing "better defenses" offers false hope to victim-survivors.

Response: We concede that no technical intervention will ever be a perfect shield. However, the goal of technical defense is not perfection, but friction. Currently, an abuser can generate AIG-NCII in minutes with little to no cost or technical skill. Adversarial defenses, even if imperfect, raise the cost of abuse by requiring knowledge to bypass. This friction may disrupt casual abuser (teenagers, ex-partners) who account for a significant volume of harassment but lack the sophistication to break these defenses. Additionally, while current defenses are brittle, this should be a call for further research, not abandonment.

AV3: The cultural and institutional incentives of AI/ML research makes engagement with AIG-NCII unrealistic. Even if researchers accept the framing of subject-centric harms, there are structural barriers to working on such topics. Lack of institutional support, concerns about the ability to publish, and reviewer discomfort with sexualized topics makes it difficult for researchers to engage with AIG-NCII.

Response: We acknowledge these incentive structures, but argue that the field cannot claim AIG-NCII is too peripheral to engage with when it is a direct downstream product of mainstream research decisions. The open datasets used to train widely-deployed diffusion models contained non-consensual images of human bodies, including CSAM (Thiel, 2023). The techniques now used for AIG-NCII (e.g., face-swapping, inpainting, fine-tuning on specific individuals) were each developed and celebrated at top venues. Having produced the capabilities, the field also bears responsibility for the mitigating their negative societal impacts. The field has shifted before. Algorithmic fairness was a fringe topic a decade ago. It took several key papers to point to the problem (Buolamwini & Gebru, 2018; Angwin et al., 2022) and efforts to form dedicated conferences (FAccT and AIES, both in 2018). The transition required individual researchers willing to legitimize the topic, conference organizers willing to allocate space, and senior scholars willing to advise students working on it. None of these moves required the field to *first* be comfortable. In fact, comfort followed the work, not the other way around. AIG-NCII can follow a similar trajectory if researchers refuse to accept "taboo" as a reason for inattention.

7. Conclusion

In this position paper, we show that there is a fundamental misalignment between the current technical interventions for deepfakes that address viewer-centric epistemic harms and the prevailing reality of subject-centric dignity harms in the form of AIG-NCII, which account for the majority of generative AI usage (Center for Countering Digital Hate, 2026; Bouchaud, 2026; Security Hero, 2023; Gibson et al., 2025). We urge the AI/ML community to realign its priorities to address these harms, else we risk exacerbating harms to victims of AIG-NCII. At the same time, we offer our call to action in the form of recommendations with a necessary constraint. Research into protections against AIG-NCII should only be undertaken when adequate safety protocols are established, including mitigating harm for researchers and establishing substantive partnerships with victim-advocates and sexual violence prevention experts. Ultimately, the research community bears a responsibility to ensure that our definitions of AI safety protect not only truth, but also the dignity of people.

Acknowledgments

We thank Amna Batool, Rosanna Bellini, and Su Lin Blodgett for conversations that helped refine the framing of this work. This material is based on works supported by the National Science Foundation under Grant 2311102.

References

How Telegram Became a Playground for Criminals, Extremists and Terrorists - The New York Times, 2024. URL https://www.nytimes.com/2024/09/07/technology/telegram-crime-terrorism.html.

Aghasanli, A., Kangin, D., and Angelov, P. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 467–474, 2023.

Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., and Emami, A. Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146:113157, 2020.

Ahmed, S., Jaźwińska, K., Ahlawat, A., Winecoff, A., and Wang, M. Building the epistemic community of ai safety. *Available at SSRN 4641526*, 2023.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

ArtificialIntelligenceAct.eu. Article 50: Transparency obligations for providers and deployers of certain ai systems. https://artificialintelligenceact.eu/article/50/, 2025. URL https://artificialintelligenceact.eu/article/50/. European Union Artificial Intelligence Act transparency obligations (Article 50).

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Bammey, Q. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

Batool, A., Naseem, M., and Toyama, K. Expanding concepts of non-consensual image-disclosure abuse: A study of ncida in pakistan. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024.

Bicker, L. Cho Ju-bin: South Korea chatroom sex abuse suspect named after outcry, March 2020. URL https://www.bbc.com/news/world-asia-52030219.

Black Forest Labs. Introducing FLUX.1 tools. https://bfl.ai/blog/24-11-21-tools, November 2024. Accessed: 2026-01-28.

Bond, S. How AI deepfakes polluted elections in 2024. *NPR*, December 2024.

Bouchaud, P. Grok Unleashed. Technical report, AI Forensics, January 2026.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Burkell, J. and Gosse, C. Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*, 2019.

Center for Countering Digital Hate. Grok floods X with sexualized images of women and children. Technical report, Center for Countering Digital Hate, January 2026.

Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., Levy, K., Dell, N., McCoy, D., and Ristenpart, T. The spyware used in intimate partner violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 441–458. IEEE, 2018.

Chen, B., Zeng, J., Yang, J., and Yang, R. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

Chesney, B. and Citron, D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

Cintaqia, P., Arya, A., Redmiles, E. M., Kumar, D., McDonald, A., and Qin, L. Stop the nonconsensual use of nude images in research. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 628–629, 2025.

Citron, D. K. Sexual privacy. *Yale LJ*, 128:1870, 2018.

CivitAI. Policy Update: Removal of Real-Person Likeness Content | Civitai, 2025. URL https://civitai.com/articles/15022/policy-update-removal-of-real-person-likeness-content.

Coalition for Content Provenance and Authenticity (C2PA). C2pa technical specification, version 2.3. https://spec.c2pa.org/specifications/specifications/2.3/index.html, 2026. URL https://spec.c2pa.org/specifications/specifications/2.3/index.html. Technical standard for Content Credentials (provenance and authenticity metadata).

Conger, K., Freedman, D., and Thompson, S. A. Musk's Chatbot Flooded X With Millions of Sexualized Images in Days, New Estimates Show. *The New York Times*, January 2026. ISSN 0362-4331.

Congress.gov. S.146 - 119th Congress (2025-2026): TAKE IT DOWN Act | Congress.gov | Library of Congress, 2025. URL https://www.congress.gov/bill/119th-congress/senate-bill/146.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Costanza-Chock, S. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

Cretu, A.-M., Kireev, K., Abdalla, A., Obinna, W., Meier, R., Bargal, S. A., Redmiles, E. M., and Troncoso, C. Evaluating concept filtering defenses against child sexual abuse material generation by text-to-image models. *arXiv preprint arXiv:2512.05707*, 2025.

Damer, N., Fang, M., Siebke, P., Kolf, J. N., Huber, M., and Boutros, F. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. *arXiv preprint arXiv:2302.01843*, 2023.

DeepMind. Synthid: A tool to watermark and identify content generated through ai. https://deepmind.google/models/synthid/, 2025. Accessed: 2026-01-XX.

Diel, A., Lalgi, T., Mellis, F. S., Teufel, A., and Bäuerle, A. The harm of deepfakes: A scoping review of deepfakes' negative effects on human mind and behavior. *AI & SOCIETY*, December 2025. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-025-02774-0.

Ding, M. L., Suresh, H., and Venkatasubramanian, S. How to stop playing whack-a-mole: Mapping the ecosystem of technologies facilitating ai-generated non-consensual intimate images. *arXiv preprint arXiv:2602.04759*, 2026.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. Leveraging frequency analysis for deep

fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., and Dell, N. "a stalker's paradise" how intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018.

Furizal, Ma'arif, A., Maghfiroh, H., Suwarno, I., Prayogi, D., Kariyamin, Lonang, S., and Sharkawy, A.-N. Social, legal, and ethical implications of AI-Generated deepfake pornography on digital platforms: A systematic literature review. *Social Sciences & Humanities Open*, 12: 101882, 2025. ISSN 2590-2911. doi: 10.1016/j.ssaho.2025.101882.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Gibson, C., Olszewski, D., Brigham, N. G., Crowder, A., Butler, K. R., Traynor, P., Redmiles, E. M., and Kohno, T. Analyzing the {AI} nudification application ecosystem. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1–20, 2025.

Gibson, J. Korea Wakes up to the Deadly Consequences of Spy Cams and Cyberbullying – The Diplomat, 2019. URL https://thediplomat.com/2019/12/korea-wakes-up-to-the-deadly-consequences-of-spy-cams-and-cyberbullying/.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guarnera, L., Giudice, O., and Battiato, S. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. In *Intelligent Systems Conference*, pp. 615–625. Springer, 2024.

Guo, Y., Qu, Z., Lu, W., and Luo, X. Anti-inpainting: A proactive defense against malicious diffusion-based inpainters under unknown conditions. *arXiv preprint arXiv:2505.13023*, 2025.

Gyevnar, B. and Kasirzadeh, A. Ai safety for everyone. *Nature Machine Intelligence*, pp. 1–12, 2025.

Hamid, Y., Elyassami, S., Gulzar, Y., Balasaraswathi, V. R., Habuza, T., and Wani, S. An improvised cnn model for fake image detection. *International Journal of Information Technology*, 15(1):5–15, 2023.

Han, C., Li, A., Kumar, D., and Durumeric, Z. Characterizing the {MrDeepFakes} sexual deepfake marketplace. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5169–5188, 2025.

Harris, K. R. Video on demand: What deepfakes do and how they harm. *Synthese*, 199(5-6):13373–13391, December 2021. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-021-03379-y.

Havron, S., Freed, D., Chatterjee, R., McCoy, D., Dell, N., and Ristenpart, T. Clinical computer security for victims of intimate partner violence. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 105–122, 2019.

Hazra, S., Majumder, B. P., and Chakrabarty, T. Ai safety should prioritize the future of work. *arXiv preprint arXiv:2504.13959*, 2025.

Henry, N. and Beard, G. Image-based sexual abuse perpetration: A scoping review. *Trauma, Violence, & Abuse*, 25(5):3981–3998, 2024.

Henry, N. and Powell, A. Sexual violence in the digital age: The scope and limits of criminal law. *Social & legal studies*, 25(4):397–418, 2016.

Henry, N., Umbach, R., Shelby, R., Beard, G., and Given, L. M. 'it's still abuse': community attitudes and perceptions on ai-generated image-based sexual abuse. *Information, Communication & Society*, pp. 1–21, 2026.

Hönig, R., Rando, J., Carlini, N., and Tramèr, F. Adversarial perturbations cannot reliably protect artists from generative ai. In *International Conference on Learning Representations*, volume 2025, pp. 70223–70263, 2025.

Hsu, C.-C., Zhuang, Y.-X., and Lee, C.-Y. Deep fake image detection based on pairwise learning. *Applied Sciences*, 10(1):370, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Ivanovska, M. and Štruc, V. Face morphing attack detection with denoising diffusion probabilistic models. *arXiv preprint arXiv:2306.15733*, 2023.

Jeon, J., Kim, W. J., Ha, S., Son, S., and Yoon, S.-e. Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. *arXiv preprint arXiv:2503.10081*, 2025.

Kang, H., Wen, S., Wen, Z., Ye, J., Li, W., Feng, P., Zhou, B., Wang, B., Lin, D., Zhang, L., et al. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*, 2025.

Keita, M., Hamidouche, W., Bougueffa Eutamene, H., Taleb-Ahmed, A., Camacho, D., and Hadid, A. Bi-lora: A vision-language approach for synthetic image detection. *Expert Systems*, 42(2):e13829, 2025.

Kim, J., Nam, Y., Kim, M., Kim, S., and Jeong, J. Blurguard: A simple approach for robustifying image protection against ai-powered editing. *Advances in Neural Information Processing Systems*, 38:28664–28706, 2026.

Lei, L., Gai, K., Yu, J., and Zhu, L. Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model. *arXiv preprint arXiv:2405.02696*, 2024.

Liao, Q., Li, Y., Wang, X., Kong, B., Zhu, B., Lyu, S., Yin, Y., Song, Q., and Wu, X. Imperceptible adversarial examples for fake image detection. *arXiv preprint arXiv:2106.01615*, 2021.

Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., and Zhao, Y. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Liu, Y., Li, Z., Backes, M., Shen, Y., and Zhang, Y. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023.

Luo, Y., Du, J., Yan, K., and Ding, S. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17006–17015, 2024.

Ma, R., Duan, J., Kong, F., Shi, X., and Xu, K. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023.

Maddocks, S. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4):415–423, 2020.

Maiberg, E. a16z-Backed AI Site Civitai Is Mostly Porn, Despite Claiming Otherwise, July 2025a. URL https://www.404media.co/a16z-backed-ai-site-civitai-is-mostly-porn-despite-claiming-otherwise/.

Maiberg, E. Hugging Face Is Hosting 5,000 Nonconsensual AI Models of Real People, July 2025b. URL https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/.

Marini, M., Ansani, A., Demichelis, A., Mancini, G., Paglieri, F., and Viola, M. Real is the new sexy: The influence of perceived realness on self-reported arousal to sexual visual stimuli. *Cognition and Emotion*, 38(3): 348–360, 2024.

Massanari, A. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346, 2017.

McGlynn, C. and Westmarland, N. Kaleidoscopic justice: Sexual violence and victim-survivors' perceptions of justice. *Social & Legal Studies*, 28(2):179–201, 2019.

Meta Platforms, I. Our approach to labeling ai-generated content and manipulated media. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/, Apr 2024. URL https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/. Meta blog post on labeling AI content across Facebook, Instagram, and Threads.

New, B., Salazar, L., Lozano, M., and Fralicks, S. Deepfakes of Elon Musk are contributing to billions of dollars in fraud losses in the U.S. - CBS Texas. https://www.cbsnews.com/texas/news/deepfakes-ai-fraud-elon-musk/, November 2024.

Nissenbaum, H. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

Ojha, U., Li, Y., and Lee, Y. J. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Olteanu, A., Barocas, S., Blodgett, S. L., Egede, L., DeVrio, A., and Cheng, M. Ai automatons: Ai systems intended to imitate humans. *arXiv preprint arXiv:2503.02250*, 2025.

Oversight Board. New decision addresses meta's rules on non-consensual deepfake intimate images. https://www.oversightboard.com/news/new-decision-addresses-metas-rules-on-non-consensual-deepfake-intimate-images/, Jul 2024. URL https://www.oversightboard.com/news/new-decision-addresses-metas-rules-on-non-consensual-deepfake-intimate-images/. Oversight Board press release on deepfake intimate image policy and Meta's rules.

Ozden, T. C., Kara, O., Akcin, O., Zaman, K., Srivastava, S., Chinchali, S. P., and Rehg, J. M. Diffvax: Optimization-free image immunization against diffusion-based editing. In *The Fourteenth International Conference on Learning Representations*, 2025.

Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., and Mazibuko, T. F. An improved dense cnn architecture for deepfake image detection. *IEEE Access*, 11:22081–22095, 2023.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

Qiwei, L., Zhang, S., Pratt, S. P., Kasper, A. T., Gilbert, E., and Schoenebeck, S. A law of one's own: The inefficacy of the dmca for non-consensual intimate media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rahman, M. A., Paul, B., Sarker, N. H., Hakim, Z. I. A., and Fattah, S. A. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2200–2204. IEEE, 2023.

Raza, A., Munir, K., and Almutairi, M. A novel deep learning approach for deepfake image detection. *Applied Sciences*, 12(19):9820, 2022.

Ricker, J., Damm, S., Holz, T., and Fischer, A. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

Ricker, J., Lukovnikov, D., and Fischer, A. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Rini, R. Deepfakes and the epistemic backstop. 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rosenthol, L. C2pa: the world's first industry standard for content provenance (conference presentation). In *Applications of Digital Image Processing XLV*, volume 12226, pp. 122260P. SPIE, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Scheuerman, M. K., Hanna, A., and Denton, R. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Schoenebeck, S., Haimson, O. L., and Nakamura, L. Drawing from justice theories to support targets of online harassment. *new media & society*, 23(5):1278–1300, 2021.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22522–22531, 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.

Security Hero. 2023 state of deepfakes. Technical report, Security Hero, 2023.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., hÉigeartaigh, S. Ó., Korinek, A., et al. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*, 2023.

Sen. Durbin, R. J. D.-I. S.3696 - 118th Congress (2023-2024): DEFIANCE Act of 2024, July 2024. URL https://www.congress.gov/bill/118th-congress/senate-bill/3696. Archive Location: 2024-01-30.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. Y. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Sinitsa, S. and Fried, O. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4067–4076, 2024.

Solaiman, I. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 111–122, 2023.

Song, H., Huang, S., Dong, Y., and Tu, W.-W. Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*, 2023.

Song, X., Guo, X., Zhang, J., Li, Q., Bai, L., Liu, X., Zhai, G., and Liu, X. On learning multi-modal forgery representation for diffusion generated video detection. *Advances in Neural Information Processing Systems*, 37:122054–122077, 2024.

Stark, L. Facial recognition, emotion and race in animated social media. *First Monday*, 2018.

Sun, J., Wang, J., Nie, W., Yu, Z., Mao, Z., and Xiao, C. A critical revisit of adversarial robustness in 3d point cloud recognition with diffusion-driven purification. In *International Conference on Machine Learning*, pp. 33100–33114. PMLR, 2023.

Sun, K., Chen, S., Yao, T., Liu, H., Sun, X., Ding, S., and Ji, R. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024.

Tenbarge, K. Grok is being used to mock and strip women in hijabs and sarees. WIRED, Jan 2026. URL https://www.wired.com/story/grok-is-being-used-to-mock-and-strip-women-in-hijabs-and-sarees/. [Online; accessed 27-Jan-2026].

Thiel, D. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23(3):131, 2023.

TikTok. About ai-generated content. https://www.tiktok.com/tns-inapp/pages/ai-generated-content, 2026. URL https://www.tiktok.com/tns-inapp/pages/ai-generated-content. TikTok support page on AI-generated content definitions and labeling requirements.

United Nations. Universal declaration of human rights, December 1948.

Van der Nagel, E. Verifying images: Deepfakes, control, and consent. *Porn Studies*, 7(4):424–429, 2020.

Van Le, T., Phung, H., Nguyen, T. H., Dao, Q., Tran, N. N., and Tran, A. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Wagner, L. and Cetinic, E. Perpetuating misogyny with generative ai: How model personalization normalizes gendered harm. *arXiv preprint arXiv:2505.04600*, 2025.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Wei, M., Yeung, C., Roesner, F., and Kohno, T. " we're utterly ill-prepared to deal with something like this": Teachers' perspectives on student generation of synthetic non-consensual explicit imagery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., and Li, W. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025.

Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Widder, D. G., Nafus, D., Dabbish, L., and Herbsleb, J. Limits and possibilities for "ethical ai" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2035–2046, 2022.

Williamson, E., Gregory, A., Abrahams, H., Aghtaie, N., Walker, S.-J., and Hester, M. Secondary trauma: Emotional safety in sensitive research. *Journal of Academic Ethics*, 18(1):55–70, 2020.

Wu, H., Zhou, J., and Zhang, S. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, 2025.

Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., and Yu, N. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

Yim, H. South Korea to criminalize watching or possessing sexually explicit deepfakes, 2024. URL https://www.reuters.com/world/asia-pacific/south-korea-criminalise-watching-or-possessing-sexually-explicit-deepfakes-2024-09-26/.

Yu, Z., Ni, J., Lin, Y., Deng, H., and Li, B. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12765–12774, 2024.

Zhang, B., Li, S., Feng, G., Qian, Z., and Zhang, X. Patch diffusion: a general module for face manipulation detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3243–3251, 2022.

Zhang, L., Liu, X., Martin, A. V., Bearfield, C. X., Brun, Y., and Guan, H. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024.

Zhang, Y. and Xu, X. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023.

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M., and Lin, M. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

Zhu, Y., Wang, X., Chen, H.-S., Salloum, R., and Kuo, C.-C. J. A-pixelhop: A green, robust and explainable fake-image detector. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8947–8951. IEEE, 2022.

Paper surveyed	Venue	Mentions AIG-NCII
Frank et al. (2020)	ICML	
Wang et al. (2023)	ICCV	
Corvi et al. (2023)	ICASSP	
Hsu et al. (2020)	Applied Sciences	✓
Ahmadi et al. (2020)	Expert Systems	
Zhao et al. (2023)	arXiv preprint	
Raza et al. (2022)	Applied Sciences	✓
Ricker et al. (2022)	arXiv preprint	
Liu et al. (2024)	CVPR	
Patel et al. (2023)	IEEE Access	✓
Yang et al. (2024)	CVPR	
Ricker et al. (2024)	CVPR	
Bammey (2023)	IEEE OJSP	
Chen et al. (2024)	ICML 2024	
Hamid et al. (2023)	Springer IJSA	
Luo et al. (2024)	CVPR 2024	
Liu et al. (2023)	arXiv preprint	
Ma et al. (2023)	arXiv preprint	
Rahman et al. (2023)	IEEE ICIP	
Damer et al. (2023)	arXiv preprint	
Yu et al. (2024)	CVPR	
Guarnera et al. (2024)	Intell. Sys. Conf.	✓
Wu et al. (2025)	arXiv preprint	
Aghasanli et al. (2023)	ICCV	
Song et al. (2023)	arXiv preprint	
Sun et al. (2024)	NeurIPS	
Lei et al. (2024)	arXiv preprint	
Sinitsa & Fried (2024)	WACV	
Song et al. (2024)	NeurIPS	
Zhu et al. (2022)	ICASSP	
Zhang et al. (2022)	AAAI	
Wen et al. (2025)	arXiv preprint	
Zhang et al. (2024)	NeurIPS	
Liao et al. (2021)	arXiv preprint	✓
Keita et al. (2025)	Expert Systems	
Kang et al. (2025)	arXiv preprint	
Zhang & Xu (2023)	arXiv preprint	
Sun et al. (2023)	ICML	
Ivanovska & Štruc (2023)	arXiv preprint	

Table 2. 39 papers surveyed, listed in order of number of citation each paper received in January 2026.