Deepfake, Real Harm: A Participatory Approach for Imagining Infrastructures to Combat Deepfake Sexual Abuse

Saetbyeol LeeYouk
MIT Media Lab
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
sbleeyuk@mit.edu

Joseph Seering
School of Computing
KAIST
Daejeon, Republic of Korea
seering@kaist.ac.kr

Abstract

With generative AI enabling easier production of sexually abusive content, deepfake sexual abuse has intensified, making anyone with visual data be a potential victim or perpetrator. Current moderation systems for non-consensual intimate imagery (NCII) are platform-centric, reactive, and poorly aligned with the workflows of real-time monitors and survivor supporters. To address this gap, we held participatory design workshops with 10 activists affiliated with victim advocacy and survivors experienced in combating deepfake sexual abuse in South Korea. Their insights revealed distinctive challenges, including ambiguity in content classification, barriers to evidence collection, and increased workloads and safety risks during monitoring. Participants suggested features for proactive protection, long-term case tracking, and cross-platform coordination, while emphasizing the need for conversations about data ownership and platform accountability. Based on these findings, we discuss design implications for system and policy that foster multi-stakeholder collaboration to prevent harm, strengthen cross-platform response, and reduce secondary trauma for activists.

CCS Concepts

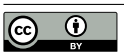
• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Keywords

deepfake, digital sexual abuse, participatory design, activism

ACM Reference Format:
Saetbyeol LeeYouk and Joseph Seering. 2026. Deepfake, Real Harm: A Participatory Approach for Imagining Infrastructures to Combat Deepfake Sexual Abuse. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3790902

CONTENT WARNING: This paper contains a discussion of sexual abuse of both adults and minors, references to non-consensual deepfake content, and descriptions of harassment that readers may find disturbing.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790902

1 INTRODUCTION

The rise of *deepfake technology* has made tools for sexual abuse fast and easily accessible. What once required technical expertise is now accomplished in minutes using app store software, Telegram chatbots [15, 92], or marketplaces trading explicit AI models and generated content [31]. Generative AI technology is a prominent tool in digital sexual violence, enabling individuals to fabricate explicit content featuring people from their personal lives. Recent estimates indicate that 98% of deepfakes shared online are sexual in nature, and 99% of explicit deepfakes depict or focus on female targets [31, 37], with reports showing a 300% increase in child-targeted sexual deepfakes in 2024 compared to 2023 [58]. In Korea, the deepfake sexual abuse cases reported to the police in 2024 increased by 128% (total 1,807 cases) compared to previous year (2023. total 793 cases) [40]. Communities on platforms like Reddit and Discord accelerate this trend by sharing prompts and techniques [31, 92]. Not stopping at simple face-swapping, deepfakes are becoming increasingly violent, degrading, and focused on dehumanizing scenarios [34, 44]. These forms of synthetic media are weaponized to blackmail, coerce, or control victims [30, 44], including in cases of sextortion and intimate partner abuse [6, 57].

This shift in scale, speed, and intent exposes the limitations of existing technical interventions for addressing non-consensual intimate imagery (NCII). Most platforms respond after harmful content is already disseminated, relying on post-hoc detection using watermarking or AI classifiers [26]. These reactive tools often fail to prevent harm or keep pace with the case volume. Although deepfakes are frequently used as a tool to facilitate or escalate physical sexual abuse, they are often deprioritized relative to other types of crimes due to limited investigative resources [61, 74].

Compounding the issue, deepfakes can be indistinguishable from other types of digital sexual abuse. In recent years, monitoring teams have both misclassified real victims as synthetic and attempted to rescue AI-generated characters [44]. In child sexual abuse material (CSAM) cases, attackers combine children's faces with adult bodies [1, 2], leading detection systems to flag the content as legal pornography and missing intervention opportunities. These challenges signal the need for new technical intervention frameworks prioritizing prevention, cross-platform coordination, and survivor-centered safety over reactive moderation.

To address this gap, this paper investigates the challenges faced by activists and support professionals working on monitoring, reporting, and responding to deepfake sexual abuse, some of whom are also survivors of deepfake sexual abuse. In South Korea, government-affiliated and civil society organizations play a central role in monitoring, evidence collection, and victim support, often stepping in

where international platforms or police responses are limited [67]. As a result, some activists become direct targets of abuse, especially when their advocacy becomes visible through news coverage or social media. Conversely, some survivors choose to become activists themselves, using their lived experience to support others. Understanding this complex overlap between professional roles and personal experience, the participants in this study were embedded within Korea's institutional framework, offering insight into both grassroots and formal intervention practices. Through participatory design, we co-imagine technological interventions that center their needs and lived experiences. This study addresses the following research questions:

- RQ1 What technical and procedural challenges do activists, and support professionals face in preventing, monitoring, and reporting sexual deepfake crimes?
- RQ2 What support needs and values should guide the design of safeguarding technologies for addressing sexual deepfakes?
- RQ3 What technical features and design strategies can improve the effectiveness and trauma-responsiveness of interventions for sexual deepfake detection and reporting?

We conducted participatory design workshops with 10 participants, including activists from nonprofit and government-affiliated victim support organizations. Three participants also disclosed having personally experienced deepfake sexual abuse. Participants shared first-hand experiences across monitoring, legal, and psychological support work. Using the Journey Mapping framework (used for defining the user journey in HCI studies) [38] participants leveraged a value sensitive design framework to identify problems and envisioned future interventions for preventing and responding to deepfake abuse [28, 75]. Through this collaborative process, we enabled participants to help shape actionable, survivor-centered strategies.

The contributions of this paper are threefold: First, we provide a detailed account of the challenges and expectations activists and support professionals face in addressing deepfake sexual abuse. Second, we identify design implications for designing safeguards that enable faster, cross-platform coordination and reduce secondary trauma for activists engaged in monitoring and reporting, highlighting the need for cross-platform, collaborative efforts among developers, platforms, law enforcement, and support organizations. Finally, we contribute a participatory design approach that integrates speculative design, foregrounding lived expertise and demonstrating how affected communities can actively shape the direction of sociotechnical interventions.

Although this study centers the South Korean context, deepfake sexual abuse is a rapidly escalating global problem [89, 91, 95]. Our focus on Korea serves two aims: to broaden trust and safety discourse beyond Western-dominated perspectives, and to highlight Korea's uniquely developed ecosystem of activists and support professionals, whose coordinated practices offer a valuable lens for informing future approaches to addressing deepfake sexual abuse.

2 RELATED WORKS

Based on the prior research and documented cases related to deepfake sexual abuse, this section outlines how sexual deepfakes present distinct challenges compared to earlier forms of non-consensual intimate image (NCII) abuse, reviews current detection and safeguarding technologies, and explores participatory and speculative design approaches for developing technology in sensitive contexts.

2.1 Distinctive Characteristics of Deepfake Sexual Abuse

Generative AI has introduced new challenges to NCII by enabling the creation of hyper-realistic sexual deepfakes without any original explicit content. While some synthetic imagery may be consensual-e.g., for gender expression or artistic use-this paper focuses on non-consensual cases. Here, harm arises from the absence of consent rather than the sexual nature of the content itself. This distinction complicates detection and moderation, as legality cannot be inferred from visual features alone. Unlike traditional NCII, which relies on the unauthorized use of real intimate content, deepfake abuse fabricates explicit media without prior sexual material. This marks a profound shift in how victimization occurs. One of its most disturbing characteristics is that *anyone* can be targeted without ever creating or sharing intimate images. Simultaneously, *anyone* with access to publicly available facial imagery can produce such content, expanding the pool of victims and potential perpetrators. [14, 49, 53, 57]

The underlying mechanisms driving this phenomenon involves open-source generative AI models, many of which are fine-tuned variants of academic research [71]. These models are available on platforms like Github, Hugging Face, and Civitai, and supported by user-friendly tutorials shared across sites like Reddit and Discord [31, 34]. In South Korea, Telegram emerged as a central hub for content creation and distribution. Bots integrated into Telegram chatrooms, often discovered via links shared on Twitter or Facebook, enable users to generate deepfakes easily [11, 39]. Alarmingly, such bots are sometimes connected to broader illicit online economies, including illegal gambling, using explicit content as bait to attract and further exploit users [11, 15, 39].

This accessibility led to a troubling demographic shift among perpetrators. According to 2024 legal data from South Korea, 80.4% of reported offenders in sexual deepfake cases were teenagers [11]. This surge reflects the growing normalization of synthetic sexual violence within youth digital cultures. Young people are consuming and creating non-consent deepfakes, targeting peers, educators, or celebrities, without fully comprehending the ethical or legal implications.

Under South Korea's Act on the Prevention of Sexual Violence and Protection of Victims,[1] the creation of sexual deepfakes with the intent to distribute constitutes a criminal offense. Moreover, even without distribution, the mere possession or viewing of such material is also punishable under the law. Despite the severe harms inflicted, deepfake sexual content exists in a legal gray area. In jurisdictions like South Korea, prosecution under current sexual violence statutes requires demonstrable evidence that the perpetrator deliberately targeted a specific individual—such as by inputting their identity into a prompt. If the accused denies using actual

[1] Act on Special Cases Concerning the Punishment, etc. of Sexual Crimes, Article 14-2. Amended October 16, 2024. This law criminalizes the editing, distribution, and possession of sexually explicit synthetic media created without consent, including deepfake content.

images or claims the generated result is coincidental, establishing criminal intent becomes challenging [19, 42].

Although legal ambiguity is still debated, the psychological consequences for victims are comparable to those of NCII abuse. Victims frequently report post-traumatic stress disorder (PTSD), depression, anxiety, and social withdrawal [7, 32, 57, 97]. Simulated abuse adds an additional layer of harm, as victims often struggle to validate their experience as a violation, despite enduring reputational, emotional, and psychological damage. Conversely, psychologists alarms that for perpetrators, repeated exposure to sexual deepfakes, particularly involving socially taboo scenarios or casting their peers, may distort their sexual perceptions and contribute to unrealistic sexual expectations [57].

As the production, sharing, and speed of deepfakes scales, law enforcement agencies are facing significant barriers in responding to monitored cases. In South Korea, the number of reported cases increased by 80% between 2021 and 2024; however, cyber investigators are responsible for hundreds of cases, limiting timely intervention [16, 69]. The encrypted messaging platforms and anonymous sharing practices further hinder identifying perpetrators and content removal. Investigators also warn of emerging complications in content monitoring. Resources may be misallocated to investigating synthetic "victims," while real victims are overlooked or misclassified as AI-generated avatars—raising serious concerns about the accuracy and ethics of current detection practices [2]

Since digital sexual abuse cases became more complex and widespread, South Korea has established a multi-layered support ecosystem, including: (1) a national Digital Sex Crime Monitoring and Response Center that coordinates policy, monitoring, and investigations across the country; (2) regional digital sexual crime support centers that work in partnership with local police; and (3) specialized digital sexual violence counselors embedded within victim support centers to provide tailored assistance for digital sex crime survivors [65, 67]. These institutions help fill the gaps between law enforcement and the Korea Communications Standards Commission, which depends on the cooperation of global platforms for content takedown. In practice, activists and support organizations play a central role in collecting evidence, identifying suspects by undercover investigation, monitoring platforms, submitting takedown requests, and supporting the mental and physical recovery of survivors - performing essential work that lies outside formal state capacity. Also, journalists played critical roles in exposing networks of crime thorugh prolong monitoring of Telegram chatrooms [60]. To hold platforms accountable, South Korea enacted Article 64-5 of the *Act on Promotion of Information and Communications Network Utilization and Information Protection*,[2], which mandates that platforms and web-hard service providers designated as NCII content distribution-prevention managers publicly disclose their annual complicance reports. These reports must detail their prevention efforts, processing outcomes for takedown requests, and the appointment and operation of responsible human resource for preventing distribution of NCII, including deepfake sexual abuse and CSAM.

2.2 Interventions to Safeguard Against Digital Sexual Crimes

Efforts to prevent and respond to digital sexual crimes include deepfake sexual abuse, span AI model moderation, content detection, investigation, and enforcement [4, 26, 51, 56, 80, 98]. However, current interventions remain insufficient. In terms of AI model governance, platforms such as Google Colab [93] restrict using their infrastructure for training deepfake models and actively flag known deepfake training libraries. GitHub similarly banned hosting sexual deepfake repositories, and OpenAI's video generation model Sora explicitly bans generation of deepfake sexual content by prohibiting creation of NCII, graphic sexual content, and the use of a person's likeness without consent; yet enforcement gaps persist, highlighting moderation blind spots [35, 66, 72]. For the open-source community ecosystem like Midjourney, Stable Diffusion, Hugging Face, and Civitai similiarly prohibit nudity, pornographic, or explicit content, recent audits show that sexually explicit deepfakes of celebrities remain easily discoverable through models that do not explicitly advertise such capabilities, revealing significant failures in model labeling and moderation. [34, 64]. For example, on the Civitai platform, with over 34,000 unique models seeking the generation of a 'Celebrity', spanning almost 15 million downloads [34]. Beyond model moderation, watermarking technologies such as Deepmark trace the origins and dissemination pathways of AI-generated content for faster takedowns [86].

Detection technologies primarily target child sexual abuse material (CSAM), relying on perceptual hashing and classifier-based models [43, 76, 88, 99]. Tools like Microsoft's PhotoDNA and NCMEC's hash databases are commonly used but lack interoperability across platforms [29, 63, 68]. Investigative systems, such as Pinterest's Guardian [22], apply rule-based and machine learning techniques to detect policy violations and trigger automated enforcement actions. However, adult-targeted sexual deepfakes are typically categorized as consensual adult content, obscuring their abusive, non-consensual nature, and highlighting a detection logic gap failing to account for this materials synthetic, harmful character.

Content moderation and enforcement rely on systems like Meta's PDQ and Hasher-Matcher-Actioner frameworks, as well as external pipelines such as NCMEC's CyberTipline [68]. Yet, the absence of feedback loops between investigators and detection systems, insufficient support for submitting contextual evidence (e.g., chat logs), and lack of protection for moderators exposed to harmful content limits enforcement [26, 29].

Outside of automated detection systems, victims frequently use platform-based reporting mechanisms to request takedowns of non-consensual content. However, these systems often place an undue burden on the victim. Platforms may require detailed descriptions of the content, legal accountability for false reports, and submission of personal identification documents [8, 33]. Victims are sometimes expected to reference specific laws to justify their takedown requests, reflecting a platform-centered rather than victim-centered approach [8, 33]. Delays in action and follow-up further compound emotional distress, eroding trust in platform accountability.

[2]Service providers meeting criteria set by Presidential Decree (e.g., user count, revenue, business type) must submit an annual transparency report to the Korea Communications Commission by January 31. The report must detail the handling of illegal content (including deepfake sexual abuse) efforts to prevent distribution, takedown requests, and designation of a distribution prevention officer.

2.3 Participatory and Speculative Design in HCI

Recent years have seen growing recognition for deeper community and end-user participation in designing and developing emerging technologies, particularly those with significant societal impact. Participatory design (PD) was widely adopted in HCI to ensure that technologies reflect their user's lived realities, needs, and values [78, 78, 81]. This participation often takes the form of stakeholder consultation during key stages of the design process, such as need-finding, scenario development, or usability testing, where users contribute feedback based on their expertise or experience [70, 79, 84].

In cybersecurity and digital harm mitigation contexts, participatory methods have gained traction as a way to uncover context-specific vulnerabilities and co-design protective interventions [45, 46, 70, 77, 77, 84]. As Bellini et al. argue, safer and more respectful research on digital safety requires that at-risk participants are given clear information and the autonomy to decide how they engage with researchers—ensuring they retain agency over their own safety [12]. For example, Zhai et al. conducted participatory workshops with elderly individuals to evaluate and critique simulated deepfake scams, helping to inform the design of fraud detection tools tailored to their demographic needs [100]. However, existing participatory approaches often fall short in engaging directly with populations deeply affected by harmful technologies—such as survivors of digital sexual violence. These individuals face heightened emotional risk, as recounting their experiences may trigger re-traumatization. Consequently, many design processes either exclude these voices or involve them in limited, extractive ways that fail to center their agency or emotional safety.

To navigate these challenges and ethically engage participants with lived experience of harm, our study draws on speculative design as both a protective and provocative method. Rather than eliciting direct recollections of trauma or prescriptive solutions, speculative design creates a fictional or future-oriented context in which participants are invited to imagine alternative futures [21, 41]. This allows for critical reflection on current limitations, ethical dilemmas, and structural inequalities embedded in technology [85, 101]. Jang et al. describe speculative design as "thinking in the future perfect tense"—a practice of imagining what will be recognized, prompting designers to reflect on the long-term societal implications of today's technological decisions [41]. In our case, speculative scenarios allowed participants to discuss systemic problems and possibilities without requiring disclosure of personal trauma, while surfacing collective visions for safer, more equitable digital infrastructures.

By combining participatory and speculative design, we aim to bridge the gap between experiential knowledge and technical development, fostering inclusive design processes that center care, agency, and imagination in the face of sensitive and emerging harms.

3 Methodology

To explore how technical interventions can better align with frontline workflows for preventing and responding to deepfake sexual abuse, we conducted participatory design workshops with 10 participants. The goal of these workshops was to (1) identify the procedural and technical challenges that activists and support professionals face across stages of prevention, monitoring, reporting, and response and (2) to co-design technical interventions by grounding design opportunities in activists and support professionals' lived experiences and imagination. Participants included activists and support professionals affiliated with victim support or monitoring organizations, three of whom were also survivors of deepfake sexual abuse, who had experience addressing deepfake sexual abuse as well as broader forms of digital sexual abuse. To ensure psychological safety, two facilitators with expertise in counseling survivors of sexual violence reviewed the workshop materials, and one of the two counselors facilitated each session. We conducted three sessions in total, each involving 3–4 participants and one facilitator.

The workshop (Figure 1) consisted of a brief **Introduction and ice-breaking** session followed by three core activities: **(1) Journey Mapping: Identifying Challenges Across Stages of Deepfake Sexual Abuse Response**, **(2) Problem Definition**, and **(3) Speculative Design: Envisioning Future Headlines Announcing the End of Deepfake Sexual Abuse**. The following sections outline our ethical considerations, participant recruitment, workshop materials, and the procedural structure of each session.

For the workshop, we adopt the legal definition of deepfake sexual crime from Article 14-2 of the Act on Special Cases Concerning the Punishment of Sexual Crimes in South Korea, which includes the non-consensual distribution or for-profit dissemination of synthetically manipulated media—such as a person's face, body, or voice—depicted in a sexually explicit manner.[3] To focus on the distinctive characteristics introduced by advances in generative AI, we limited our scope to cases involving AI-generated or AI-assisted synthetic media. Manipulations created solely using basic photo editing tools (e.g., Photoshop) were excluded, as they do not reflect the technical and sociocultural shifts of crime that are brought by generative AI.

3.1 Ethical Considerations

Given the sensitive nature of the topic, particularly in relation to participants' psychological well-being and the protection of personal information, this study adopted multiple strategies to ensure ethical integrity and participant safety. The study procedure was approved through our institution's Institutional Review Board (IRB).

To prevent the risk of secondary victimization through workshop activities or materials, all recruitment forms, pre-survey questionnaires, and workshop materials were reviewed and revised in consultation with facilitators who possess professional expertise in counseling survivors of sexual violence at victim support organizations in South Korea. Each workshop was conducted with this trained facilitator present to guide discussions with sensitivity and ensure participant anonymity and emotional safety throughout the session.

Participants were clearly informed of their rights and the voluntary nature of the study. Prior to signing the consent form, participants were informed that they were free to disclose as much or as little personal experience as they felt comfortable with and

[3] Act on Special Cases Concerning the Punishment, etc. of Sexual Crimes, Article 14-2. Amended October 16, 2024. This law criminalizes the editing, distribution, and possession of sexually explicit synthetic media created without consent, including deepfake content.

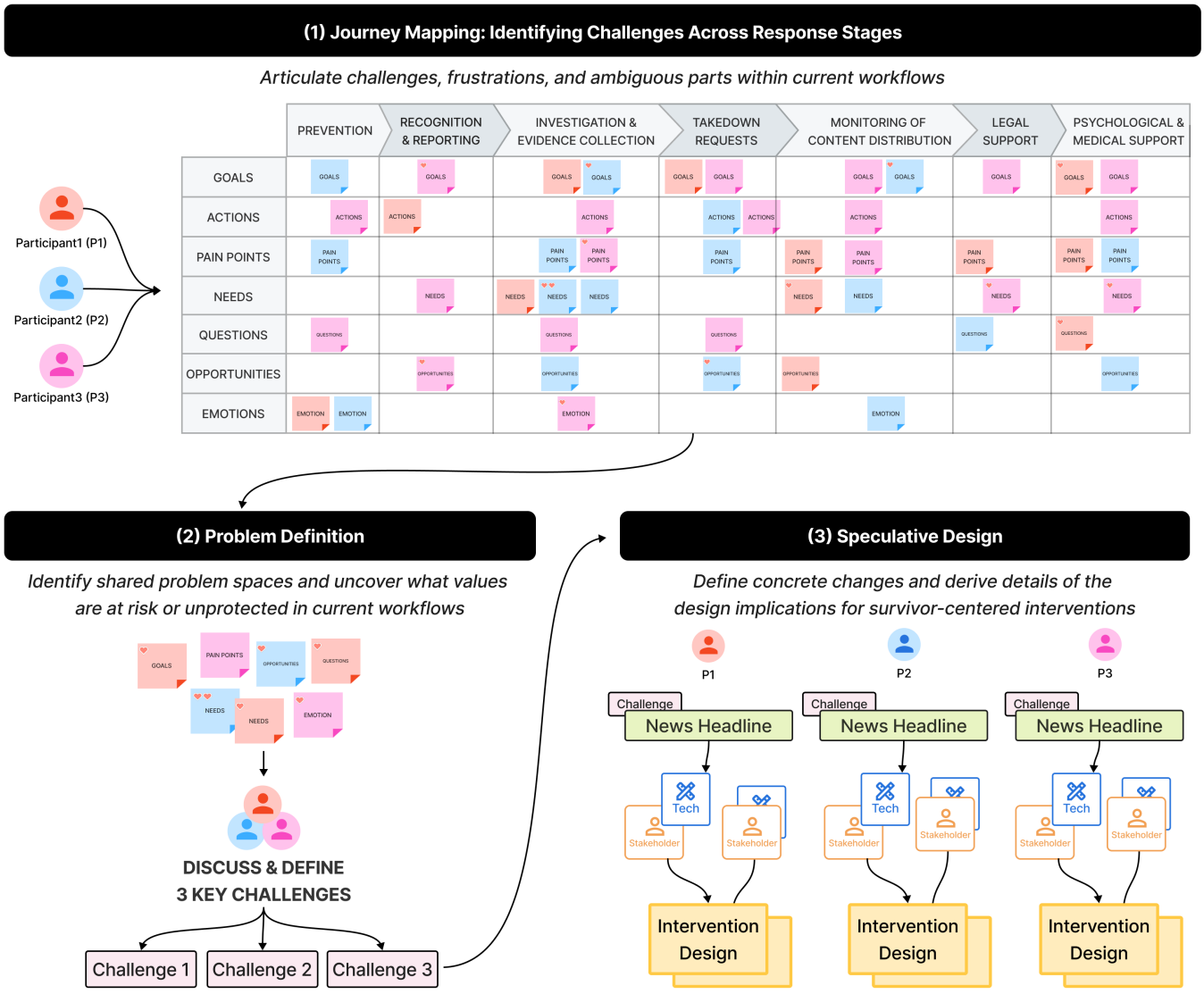


Figure 1: Illustration of the participatory design workshop process. Participants completed an onboarding session followed by three activities on a Miro board. Activity 1 was conducted individually, while Activities 2 and 3 combined individual ideation and group discussion.

could choose not to respond to any questions that felt burdensome. They were given the option to join the workshop using a pseudonym and to keep their camera turned off while attending via Zoom. At the beginning of each session, participants were encouraged to notify the researchers or facilitators if they experienced any discomfort or found the content distressing. The rules for confidentiality were also clarified in the beginning of the workshop, stating that the discussion inside the workshop may not be shared outside the group. Moreover, participants were reminded to avoid sharing sensitive information about about victims that they supported in their professional or activist capacity that could be used to identify those victims. They were also assured that they could withdraw or discontinue their participation at any time without providing a reason.

3.2 Recruitment

The site of study for this work was South Korea. As noted above, we focused on the Korean context in part because conversations in trust and safety have traditionally been biased toward Western voices and in part because of the value of understanding the unique infrastructures that exist in South Korea to combat sexual violence and exploitation. All of the participants and facilitators were based in South Korea. We recruited participants who met at least one of the following criteria: (1) survivors of deepfake sexual abuse

over the age of 19,[4] (2) individuals with experience in monitoring or supporting takedown efforts for deepfake and digital sexual abuse content (e.g., staff at national or local victim support centers, law enforcement officers), (3) professionals who had provided psychological, legal, or medical support to survivors (e.g., counselors, legal advocates), or (4) journalists who had conducted undercover investigations or reported on deepfake crimes through survivor interviews and field reporting.

The recruitment survey allowed applicants to select their status from three categories: (1) survivor over age 19, (2) survivor and activist, or (3) activist. For categories (2) and (3), "activist" was defined as individuals with at least one year of experience in monitoring, supporting, or reporting deepfake or broader digital sexual abuse. The eligibility criteria and affiliation categories were adapted from the Seoul Digital Sex Crime Victim Support Center's guidelines and reflect the operational structure of South Korea's activist networks [67].

Recruitment materials were distributed via (1) an online bulletin board hosted by a university's Center for Human Rights and Gender Equality, (2) social media platforms (e.g., Instagram, university online communities), and (3) email outreach to activist and survivor support organizations. Snowball sampling was also encouraged via participant referrals.

The recruitment survey (See Appendix A) collected different information depending on category. For survivors and activist-survivors, the form asked about willingness to share experiences, type of deepfake sexual crime experienced, their current response to the incident, challenges in recognizing and managing the crime. For activists and activist-survivors, the survey asked for affiliation, duration of service, number of survivors they've supported, specific tasks (e.g., monitoring, legal aid), and familiarity with technical interventions. For activist-survivor participants—those who both experienced abuse and support other survivors—we asked them to respond to questions for both survivors and activists to capture the dual nature of their experiences. A follow-up pre-survey (See Appendix B) for selected participants assessed expertise across intervention stages, specific pain points, and preferences for anonymity. Participants received 60,000 KRW (approx. 43 USD) in compensation for participation.

3.2.1 Facilitators. Facilitators were selected based on one of two criteria: (1) holding a master's degree or higher in psychology or women's and gender studies, or a related field with experience in victim-centered counseling or research; or (2) having at least one year of counseling experience with survivors of sexual violence at a human rights or victim support institution. Applicants completed a pre-survey detailing their current affiliation, counseling focus, and work with sexual crime survivors. A curriculum vitae was also required for qualification verification.

Selected facilitators were compensated 250,000 KRW (approx. 180 USD) for facilitating two workshops and reviewing the safety of workshop materials and analytical results. During sessions, facilitators ensured trauma-informed engagement and emotional safety, while the research team guided the discussion and workshop flow according to protocol.

[4]In South Korea, 19 is the age of majority.

3.3 Participants

A total of 12 participants were invited based on their direct experience addressing deepfake sexual abuse. Individuals whose expertise focused solely on other forms of digital sexual abuse were not included. Out of the 12 participants, 10 accepted and participated in the workshops, along with two trained facilitators. (see Table 1) Given the sensitivity of the topic and the potential risk of disclosing confidential case information, some participants were required to obtain approval from their affiliated organizations prior to participation. These constraints posed challenges for recruitment and limited the pool of eligible participants. Nevertheless, due to the highly specialized nature of participants' experiences and the in-depth procedures of the workshop, we were able to draw meaningful insights even from a smaller sample size than is often used in other research methods. We discuss this further in the limitations section.

Each workshop session included 3–4 participants and one facilitator. Three participants identified as survivors of deepfake sexual abuse. While we initially aimed to recruit participants who were primarily survivors, our final participants included only activist-survivors and activists. This may reflect the sensitive nature of deepfake abuse, where those willing to publicly share their experiences are already engaged in advocacy or peer support roles. During the workshop, when activist or professional participants shared their personal experiences, facilitators guided them to confirm whether they were discussing their own experience as a survivor or representing other survivors that they've supported. Activist-survivors typically specified, 'It's from my previous experience that...' to distinguishing their own experiences, feelings, responses from the experiences they had supporting victims as an activist. Participants had varied technological literacy regarding moderation tools for generative AI, social media platform moderation tools, and illegal content reporting procedures, shaped by their professional roles. To organize each workshop session, participants were grouped based on shared affiliations and primary areas of experience. This grouping allowed participants to engage more openly with others who shared similar organizational roles and responsibilities, and reduced the risk of conflict or hesitancy that might arise in mixed-stakeholder discussions. Workshop 1 (W1) consisted of counselors from sexual violence centers with extensive experience supporting victims throughout the entire process—from initial recognition and police reporting to takedown requests and psychological or medical assistance. Workshop 2 (W2) included activists from specialized digital sexual crime organizations. While some NGOs and state-operated institutions are permitted to report direct takedown requests to platforms under Korea's Telecommunications Business Act, participants in W2 are not the part of these institutions and therefore lack formal authorization. As such, platforms often require additional verification to confirm they are acting on behalf of the victim. Instead, they play a key intermediary role by manually monitoring and collecting evidence of distributed content, helping survivors compile documentation necessary for filing formal police reports. As such, their technical literacy tended to be moderate. Workshop 3 (W3) brought together participants with experience in investigative journalism or government-affiliated monitoring organizations who had conducted undercover investigations on

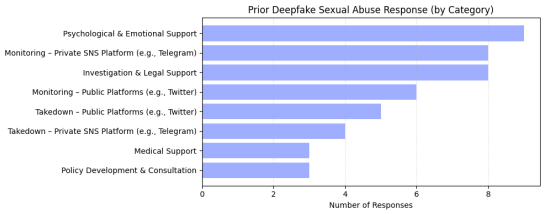


Figure 2: Participants' prior experience across domains related to digital sexual abuse response, such as legal advocacy, mental health support, and monitoring.

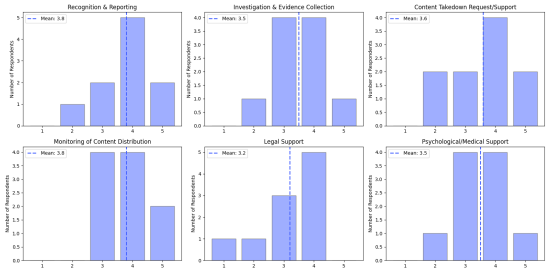


Figure 3: Participants' self-assessed expertise (rated 1 to 5) across various stages of deepfake sexual abuse response, including detection, reporting, and survivor care.

Telegram and participated in content removal processes. This group demonstrated the highest technical proficiency. While this grouping supported in-depth sharing, we acknowledge that it limited opportunities for direct cross-stakeholder interaction.

Prior experience was most concentrated in (1) psychological and emotional victim support, followed by (2) content monitoring on private platforms (e.g., Telegram), and (3) investigation and legal support (see Figure 2). As illustrated in Figure 3, participants rated themselves most knowledgeable in monitoring the distribution of deepfake content, and least confident in legal support processes. Notable variation in self-assessed expertise for investigation and takedown indicated differing professional backgrounds and access to institutional resources.

3.4 Workshop Protocol

Three participatory design workshops were conducted remotely via Zoom[5] using Miro Board.[6] Prior to the sessions, facilitators attended a *pre-training session* to review the protocol, discuss safety considerations, and iterate on workshop materials. Given the open-ended nature of the output, we aligned expectations around the scope and goals of the sessions. The full structure of the workshop is illustrated in Figure 1. Prior to the workshop, participants received instructions on how to use the Miro board, a definition of deepfake sexual abuse to help bound the discussion, and an overview of the workshop structure.

[5]https://zoom.com/
[6]https://miro.com/

3.4.1 Onboarding and Icebreaking. The workshop began with an introduction to the research objectives, workshop goals, and the overall process. To establish rapport and ensure emotional readiness, participants shared their current emotional state, motivations and expectations for participation, and areas of interest from the pre-survey regarding deepfake sexual abuse response. For participants unfamiliar with Miro, we included a brief orientation session to practice key features used during the workshop - such as adding memos and placing stickers.

3.4.2 Journey Mapping: Identifying Challenges Across Stages of Deepfake Sexual Abuse Response. Participants reflected on personal experiences across seven response stages as defined by the Seoul Digital Sex Crime Victim Support Center [67]. To support reflection, we used a virtual user journey map template [38], a common HCI tool for visualizing how goals, needs, actions, and emotional burdens evolve over time. This enabled participants to identify procedural obstacles and emotional shifts across response stages. Using this framework, participants documented each stage's *goals, actions, needs, pain points, questions, emotions, and opportunities*, guided by short prompts (e.g., "What barriers did you face?"). Below is an example from W3 P7 on *monitoring of content distribution*:

- **Goal:** "Track all abusive content."
- **Actions:** "Capture screenshots and URLs."
- **Needs:** "Technology to identify whether content is a deepfake."
- **Pain Points:** "Must blend into misogynistic chat groups to maintain access."
- **Questions:** "If perpetrators claim the image only 'resembles' the victim, is it still actionable?"
- **Emotions:** "Secondary trauma."
- **Opportunities:** "Deepfake communities overlap with illegal gambling sites—joint interventions are needed."

Participants then responded to each other's memos with virtual stickers indicating "I agree" or "I have a question on this." Facilitators summarized high-engagement memos to lead group discussions on shared challenges, their root causes, and emotional impacts.

3.4.3 Problem Definition: Connecting Insights. Participants collaboratively defined three core problems drawn from the Journey Mapping activity.We provided a structured problem-definition template: *As a [role], I aim to do [goal], but [challenge], because [reason].* An example from W1's *monitoring* stage:

- ***As** a monitoring activist, **I aim to** prevent re-circulation of deepfake content, **but** global sites often ignore our requests **because** many do not process takedown reports for illegal material.*

To support collaborative discussion, we introduced the Value Sensitive Design (VSD) framework [28, 75]. VSD integrates ethical, social, and technical values into design and helps identify value tensions within sociotechnical systems. Thirteen VSD values were provided as conceptual prompts for grouping challenges (see Appendix C). Facilitators encouraged participants to select and justify relevant value tags, creating a shared vocabulary for articulating problem statements. We note that VSD served as a conceptual aid rather than a full VSD implementation.

Session	ID	Affiliation	Sector	Deepfake Abuse Supports	Digital Sexual Abuse Supports	Tech Literacy
W1	P1	Activist	Sexual Violence Counseling Center	1-5	5–10	Low
	P2	Activist/Survivor	NGO	10–30	10–30	High
	P3	Activist/Survivor	Sexual Violence Counseling Center	5–10	1-5	Low
W2	P4	Activist	Digital Sexual Violence Center	5–10	>30	High
	P5	Activist/Survivor	Independent Activist	5–10	1-5	Low
	P6	Activist	Digital Sexual Violence Center	1-5	10–30	Moderate
W3	P7	Government Associate	National Center (Monitoring, Takedown)	>30	>30	Moderate
	P8	Journalist	Press	1-5	1-5	Moderate
	P9	Journalist	Press	1-5	1-5	Moderate
	P10	Activist	NGO (Monitoring, Takedown, Victim Support)	>30	>30	High

Table 1: Participants' prior experience, affiliations, and technology literacy related to deepfake sexual abuse."Deepfake abuse supports" refers to the number of victims the participant has supported in cases involving AI-generated non-consensual sexual imagery. "Digital sexual abuse supports" refers to broader NCII and CSAM cases, excluding deepfake-related supports. Each number indicates an approximate range of victims the participant has supported.

3.4.4 Speculative Design: Envisioning Future Headlines Announcing the End of Deepfake Sexual Abuse. To accommodate participants with diverse technical backgrounds, we used speculative design methods to prompt creativity. First, participants imagined a one-line news headline announcing the end of deepfake sexual abuse, specifying who enacted the change, how long it took, and what outcomes were achieved. These visions served as starting points for intervention ideation. Below is an example of a speculative news headline intervention from W2:

- **Headline:** *Deepfake sexual abuse victims: Strengthened safety and reporting infrastructure to be introduced.*
- **Stakeholders Involved:** *Government, AI companies and researchers, civic communities*
- **Timeline:** *2 years*
- **Envisioned Outcome:** *Anonymous identity protection during reporting, encryption of victims' personal information*

Next, participants developed step-by-step scenarios toward achieving their envisioned futures. We provided stakeholder and technology cards to scaffold this process (see Appendix D). Stakeholder cards were selected based on stakeholder groups identified in a policy forum on deepfake sexual abuse prevention and response, hosted by the National Assembly of South Korea [9] Stakeholder cards included: (1) victim advocacy, (2) social media platforms, (3) AI companies and developers, (4) open-source AI platforms, (5) police investigators, (6) government, and (7) education. Technology cards were developed by reviewing safeguarding technologies referenced in existing NCII and CSAM prevention efforts, as discussed in Section 2.2. We curated 11 cards spanning three domains: (1) AI model management, (2) platform moderation, and (3) criminal investigation of digital sexual abuse.

Participants combined stakeholder cards and technology cards to imagine technical interventions, policies, or systems—focusing on survivor-centered needs. Importantly, participants were explicitly encouraged to avoid concerns about technical feasibility and instead focus on articulating interventions that reflected their needs, lived experiences, and survivor-centered values. This approach ensured that proposed ideas were grounded in the realities of harm and response, rather than constrained by existing technological limitations. Researchers provided clarifications during ideation. Final outputs included envisioned features for proactive protection, coordinated reporting, policy infrastructure, and educational approaches to preventing deepfake sexual abuse. Below is an example of a speculative intervention scenario from W2:

- *AI-Based Integrated Case Management System*
- **Selected Stakeholders:** *Police investigators, Government*
- **Selected Technologies:** *Privacy-Protecting Tech for Victims, Cross-Platform Reporting, Distributed Monitoring System*

3.5 Data Analysis

Each workshop session lasted an average of 123 minutes (min = 117, max = 132), and all sessions were recorded via Zoom. Recordings were transcribed, and the workshop outputs captured on the Miro board were documented using Google Docs.[7]

We conducted reflexive thematic analysis following Braun and Clarke's methodology [13], with data collection and analysis occurring iteratively throughout the study. The analysis was led by the first author, with workshop facilitators providing ongoing input to support cultural and contextual interpretation and to refine the codebook.

For analysis, the first author annotated transcripts and participant notes, using analytic memos to document emerging patterns. Facilitators reviewed annotations to ensure contextual accuracy. Based on this, the first author developed an initial codebook into two primary categories: (1) challenges and (2) speculative design

[7] https://docs.google.com/document/

ideas. For the challenges, we clustered responses based on the following chunks: insights into the ecosystem of deepfake sexual abuse, participants' goals at each response stage, barriers to achieving those goals, emotional struggles; and friction encountered when collaborating with external stakeholders such as police or social media platforms. For the (2) speculative design ideas, we organized responses into following chunks: intended goal of the proposed news headline and intervention, the key actors involved in the imagined change, the envisioned process of the intervention, and deeper societal implications underlying the proposed news headline and the intervention design. This codebook was refined as a result of a review from facilitators, who suggested refinements and modifications based on their contextual understanding of the workshops. (see Appendix D).

We then analyzed the thematic linkages between challenge categories and speculative ideas within each workshop, in order to understand how participants' problem framings shaped their envisioned interventions. From there, we iteratively grouped related codes into preliminary themes, revisiting and reclassifying subsets as needed to enhance coherence. We then refined and labeled the final themes and subthemes.

For challenge-related themes, although the participant number was limited (n=10), we observed signs of conceptual saturation—key themes emerged consistently across sessions, with limited introduction of new concepts over time, suggesting sufficient depth for qualitative insight.

Importantly, for the speculative deign themes, we did not exclude ideas that fell outside the technical interventions—such as suggestions related to comprehensive sexuality education—recognizing their relevance in a survivor-centered response framework. All participants quotes presented in this paper were translated into English with careful attention to preserving their original nuance and meaning.

4 Findings: Challenges and Design Suggestions

We present findings from the workshops through three high-level themes (Figure 4): (1) The complexity of combating deepfake sexual abuse; (2) The need for proactive platform engagement; and (3) Supporting activists against secondary trauma. Within each theme, participants identified systemic challenges, unprotected values, and proposed survivor-centered interventions grounded in their lived experience and advocacy work. From these discussions, we derived five concrete design suggestions. Building on these insights, we conclude this section with a multi-stakeholder participatory design scenario that maps these suggestions onto real-world response workflows.

4.1 Complexity in combating deepfake sexual abuse

Participants emphasized that deepfake sexual abuse is not an entirely new form of sexual crime, but rather a more sophisticated and accelerated iteration of pre-existing offenses. Prior to the widespread use of generative AI, they encountered sexually manipulated synthetic images created using ID photos or publicly shared images from victims' social media profiles. With the advancement of AI, several participants (P1, P3, P4, P6) initially hoped the new technologies would be leveraged to support prevention, monitoring, and takedown of abusive content. Instead, they expressed frustration that the technology has primarily been used to scale up the harm. The fact that cutting-edge generative AI is now at the core of these abuses has left activists feeling overwhelmed and dis-empowered. As P1 described:

> *While supporting a survivor of deepfake sexual abuse, I felt a unique sense of helplessness that was different from previous digital sexual crimes. There's nothing the survivor could have done to prevent it, and even after recognizing the incident, there was little we could do in response. Both the survivor and I felt overpowered by the technology itself. (P1)*

This section explores how deepfake sexual abuse introduces distinct procedural and emotional challenges compared to earlier forms of digital sexual violence.

4.1.1 ***Anyone Can Be a Perpetrator.*** Unlike traditional digital sexual crimes, such as hidden camera recordings or sextortion, deepfake sexual abuse eliminates many of the physical processes underlying abuse. As P6 explained, *"Earlier digital sexual crimes involved many variables that could cause the crime to fail."* For example, perpetrators had to observe victims, groom or coerce them into sexual actions, or use substances to enable the abuse which required time, effort, and interaction. These steps allowed more opportunities for victims to sense danger and seek help, such as contacting law enforcement. In contrast, deepfake sexual abuse can occur instantly without any connection with the victim. As P2 stated, *"It can happen in a single click—in just three seconds."*

The ease of execution has lowered the threshold for perpetration, increasing the number of cases. Participants noted a rise in both teenage victims and perpetrators in South Korea. Because generative models that allow Not Safe For Work (NSFW) outputs are often hosted outside regulated platforms, perpetrators gather on Discord to exchange prompt engineering tips and platform links. Also in contrast to previous digital sexual crimes, participants noted that the widespread availability and ease of use of deepfake tools appears to have significantly lowered the average age of perpetrators. P5 shared a striking example where she encountered a 12-year-old managing a Discord channel that distributed deepfake sexual abuse content during an undercover investigation.

Deepfake crimes also lack identifiable traces of the perpetrator. While the victim's face is present, the perpetrator is more likely to remain completely anonymous. P5 emphasized that when perpetrators use public profile or SNS photos, *"Anyone with internet access could create deepfake sexual content—making it nearly impossible to identify the perpetrator."* Without identifiable perpetrators, survivors and their acquaintances often enter Telegram channels themselves to monitor and collect evidence—an emotionally taxing and unsafe task. This opacity undermines the potential for justice and recovery. As P1 expressed:

> *Recovery depends on the perpetrator facing consequences and fully deleting the content. But deepfake abuse seeps into daily life differently from physical violence. Survivors live in constant fear of re-distribution [of the*

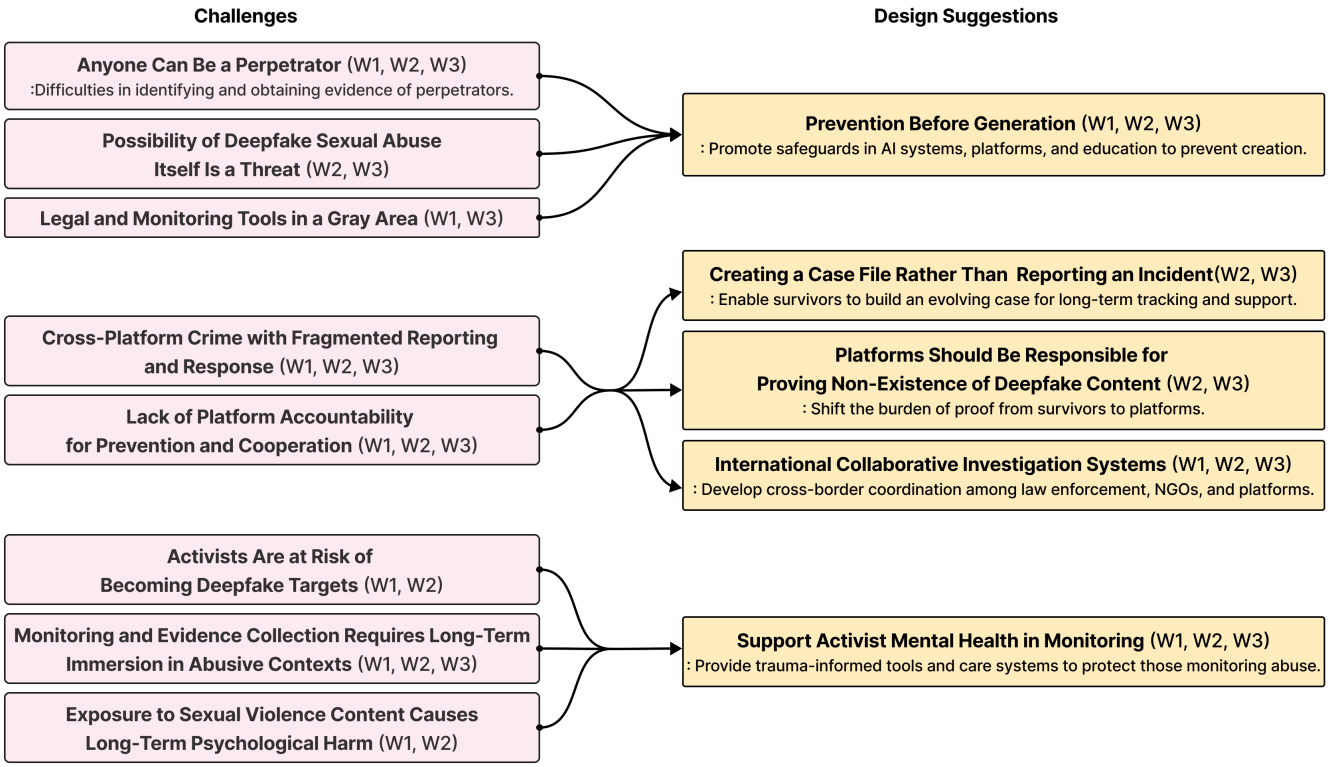


Figure 4: Overview of identified challenges and corresponding design suggestions from each workshop session. Each session's participants resulted in three to six technical intervention scenarios, to address their defined challenges and protect the values they found most at risk.

> *deepfake content] and doubt whether the content will ever truly disappear [...] That makes them lose any trust that they can fully recover under the current system. (P1)*

Perpetrators also seem encouraged by the low risk of detection. P7, who monitors Telegram channels in the government affiliated monitoring organization, noted, *"They confidently say things like 'I use a virtual phone number—they'll never catch me.' It's infuriating."*

This dissociation between process and consequence of abuse adds a new challenge that recognition of victimization is only available through third parties. Most participants (P2, P3, P4, P6, P9, P10) shared that survivors usually become aware of the content when a friend or acquaintance notifies them. P2 recalled her experience of learning about her victimization through an estranged friend who shared a link to the content and asked, "Is this really you?" Since then, she feels panic anytime someone she hasn't spoken to recently reaches out. In other cases, perpetrators themselves impersonate helpers by sending Instagram DMs pretending to inform the victim, only to screenshot the conversation and humiliate them in Telegram groups (P2). This process of identifying whether the content features them can cause further trauma and confusion.

Adding to the harm, Telegram deepfake chat rooms in Korea often share victims' personal information—such as school, workplace, or hometown—alongside the content. Channels are frequently organized by region or institution [69]. When one person is victimized, multiple others are often targeted in the same chat rooms, and this characteristic add the emotional burden for survivors. As P3 said:

> *When survivor found out she wasn't the only victim, she agonized over whether to tell the other friends who are her classmates. She feared they would go through the same pain she did. (P3)*

These findings reveal how the technical accessibility, anonymity, and viral structure of deepfake sexual abuse deepen the psychological and investigative burden on survivors, complicating not only their recovery but their ability to even detect and define what has happened to them.

4.1.2 ***The Possibility of Deepfake Sexual Abuse Itself Is a Threat.*** Participants emphasized that deepfake sexual abuse is not harmful solely through its occurrence but also through its persistent potential. Because perpetrators manipulate publicly shared social media images to generate abusive content, many individuals have reacted by setting their profiles to private or removing profile photos entirely from the social media platform. However, these reactions do not guarantee safety. Workshop participants explained that perpetrators use captured images from others' social media

or crop victims' faces from group photos, making self-protection nearly impossible.

This foundational characteristic renders the crime unpreventable on the individual level. Multiple participants (P1, P2, P3, P4, P6, P8) noted that *"anyone can be a victim"* and *"anyone around them can access generative AI tools with ease"*, creating a climate where the mere possibility of abuse threatens social trust and psychological stability. In response to incidents involving deepfake generation using school graduation photos, one elementary school in Korea required parents to submit a consent form acknowledging legal risks before distributing yearbooks [48].

P3, who supported a case involving high school students, described:

> *A male student created deepfake pornography using photos taken during a class trip. The victims were all the female students in the class. It was an act that abandoned not only courtesy or technical responsibility but fundamental human respect. (P3)*

Participants expressed concern regarding the rise of teenage perpetrators and victims, noting that this trend has not only increased fear but is also being used to regulate the behavior of teenage girls. P8 stated:

> *There were cases where male students secretly took photos of girls in their class and created deepfakes to share among peers. If a girl spoke out about feminism, the boys would tease, 'Let's make a deepfake of her.' It made girls hesitant to express their opinions, self-censoring, 'Could I become a deepfake victim just for speaking up?' (P8)*

Participants further noted that even when warning signs appear within school communities, the absence of explicit evidence makes it difficult to initiate any response. P1, P2, P4, P6, and P8 collectively highlighted how the widespread availability of generative tools has normalized the threat, making it difficult to differentiate between actual offenses and implied danger.

4.1.3 ***The Gray Area of Combating.*** The lack of reliable tools to identify deepfakes complicates monitoring efforts. Participants with monitoring experience (P2, P4, P7, P10) consistently emphasized the urgent need for AI tools to verify whether content is deepfaked. This stems from the high realism of generated images, which hinders accurate judgment by human moderators. For example, P10 shared difficulties monitoring child sexual abuse content:

> *When an adult's naked body is deepfaked with a child's face, it's incredibly difficult for moderators to determine that the victim is actually a child just by looking at the final content. (P10)*

P7 also highlighted the difficulty of recognizing celebrity deepfakes:

> *For celebrity deepfakes, moderators must know the faces of countless celebrities. But when a video is uploaded without context, it's hard to tell if it's a celebrity deepfake or general adult video, so accurately identifying and removing every deepfake is nearly impossible only with the manual monitoring. (P7)*

They further noted that the Korea Communications Standards Commission (KSCS) applies different content moderation standards for deepfake synthetic media and general explicit content. Due to this distinction, misjudgments during the monitoring process could result in missing victims, especially for child victims, who require urgent support.

Moreover, legal ambiguity further complicates evidence collection. Deepfake cases often fall into legal gray zones, especially when perpetrators deny intentional targeting. P7 said *"When a perpetrator says, 'It just looks like her. I didn't actually used her photo for generation', there's no realistic way to prove otherwise. That legal ambiguity puts us in a difficult position."* To overcome this, activists (P2, P4, P5, P10) shared that they must capture not only the deepfake content itself but also the broader digital traces—such as Telegram messages or Twitter comments—to establish the context of the crime and prove malicious intent.

4.1.4 ***Design Suggestion 1: Prevention before Generation.*** Given that most deepfake sexual abuse content is created without the depicted person's consent, participants emphasized the need for interventions that prevent harm before content is generated or shared.

To limit generation, participants suggested mandating restrictions on harmful prompts at the input stage of generative AI systems. For instance, platforms could block the use of specific terms such as "deepfake" or "nude" when users attempt to register them as chatbot descriptions (P6). Similarly, P2 and P4 proposed regulating search and naming functions so that AI models or chatrooms promoting deepfake tools cannot be easily discovered or shared via platform search functions. To prevent dissemination, P4 envisioned pre-screening mechanisms that would temporarily block uploads suspected of containing deepfake sexual content—using AI to flag content upon upload and delay public visibility until reviewed.

However, distinguishing between non-consensual and consensual synthetic sexual content remains a complex challenge. Outside of the Korean context, Pornhub began regulating deepfakes in 2018 by requiring uploader identification and manual review prior to publication [73]. This process allows individuals to upload consensual deepfakes of themselves while attempting to block non-consensual content. Yet, enforcement gaps remain—many platforms still allow searches for "deepfake" content featuring celebrity names are still readily possible [96]. This raises questions about the sufficiency of current identity verification processes and the extent to which consent is meaningfully validated.

To address this, platforms must go beyond develop mechanisms that verify the consent of depicted individuals at upload. While some technical infrastructures used for CSAM detection, such as automated scanning and hashing, could be adapted to address deepfake sexual abuse, the nuances of adult consent demand tailored approaches. To adapt this context, regulation must also address the open distribution of generative models, rather than focusing solely on content dissemination. Will et al. [34] note that platforms like Hugging Face and Civitai, despite prohibiting non-consensual sexual content in their Terms of Service, continue to host models capable of generating sexualized images of identifiable individuals, especially celebrities. Regulating model development and distribution—especially when targeting real people—may more effectively curb the production and circulation of non-consensual deepfakes.

As highlighted in Thorn's "Safety by Design for Generative AI" report [90], cross-sector collaboration among NGOs, government agencies, and technology platforms is essential for developing effective governance approaches that address both harmful outputs and the models that generate them. This includes building detection systems, restricting model access, and enforcing accountability mechanisms. However, implementing such frameworks requires careful cultural adaptation; for instance, in South Korea, even consensual adult content is subject to stricter legal and platform regulation than in many Western contexts, which may lead to differing thresholds and norms for content moderation and removal.

Beyond technical controls, participants also stressed the importance of societal interventions. They called for ethical AI education and comprehensive sex education to be integrated as long-term preventative strategies to address the conditions that enable the misuse of generative technologies (P1, P2, P3, P4, P6, P9, P10). Specifically, participants highlighted that educational programs must correctly frame deepfake sexual abuse not as a generalized form of cyberbullying, but as a manifestation of systemic gender-based violence rooted in the objectification of women (P10). They emphasized the need to address how such abuse reshapes school dynamics, instills fear among students, and contributes to a culture of surveillance and humiliation.

However, participants also noted that awareness and understanding of deepfake abuse vary widely among teachers and administrators. Thus, educational materials should be co-developed with gender equity experts, reviewed by interdisciplinary associations, and implemented through top-down mandates from the Ministry of Education (P2, P7).

4.2 The Need for Proactive Platform Engagement

Current platform responses to deepfake sexual abuse remain largely reactive, placing the burden on victims to detect, report, and prove harm. Participants emphasized the need for proactive platform engagement, calling for accountability through transparent action and survivor-centered tools. This section outlines design directions including cross-platform case tracking, systems that shift the burden of proof away from victims, and long-term monitoring features that update survivors on takedown progress. Participants envisioned platforms not as passive intermediaries, but as active agents in preventing harm and supporting recovery.

4.2.1 ***Cross-Platform Crime, Fragmented Report and Response.*** Deepfake sexual abuse is not an isolated incident, but rather part of a **broader networked system of abuse that spans multiple platforms.** Participants emphasized that perpetrator communities operate with structured hierarchies, where deepfake content is exchanged as currency that can be paid toward granting access to higher-quality models or more explicit material. This ecosystem spans multiple services: participants explained that Instagram and messenger apps are used to collect victims' personal images (P2, P3, P4, P7, P10), Discord to share prompt engineering tips and model usage guides (P5, P8, P9), and X (formerly Twitter) or Threads to promote deepfake tools or Telegram links (P4, P5, P6, P10). In many cases, perpetrators use public SNS photos to solicit deepfake creation requests by posting messages like "Make a deepfake of me with this photo." Telegram then serves as the main platform for storing and sharing the abusive content, which is again advertised through X or Discord using deepfake samples as promotional flyer.

Due to this complex and collaborative ecosystem, monitoring becomes a labor-intensive task. As P4 explained, *"To verify whether the same perpetrator is involved, I have to check whether a Telegram ID matches a similar X username, then go through that account's posts and comments—one by one—to track down content."* Participants described this cross-referencing process across multiple platforms as time-consuming and overwhelming (P4, P5, P6, P7, P10).

Despite the cross-platform nature of the abuse, the reporting and takedown processes remain fragmented. Each platform operates its own removal policy, provides different levels of cooperation, and responds with varying timelines and formats. Participants noted that many platforms place the burden of proof on survivors, requiring them to "prove" they are the person depicted in the content—often by submitting a government-issued ID. P2 explained,

> *Even if a site says they will take it down, survivors face a dilemma: they don't trust the site administrators, but they also want to stop the spread—so they send their ID photo despite the risks. (P2)*

4.2.2 ***Lack of Platform Accountability on Prevention and Cooperation.*** Beyond inconsistent reporting mechanisms, participants described a broader lack of transparency and accountability in how platforms engage with law enforcement. As P1 explained, *"Platforms claim to cooperate with law enforcement, but there's no way to verify what they've actually done or whether they took any action on reported content."* (P1) In response to this, Korea introduced a requirement in 2022 for platforms and web-hard service providers designated as NCII distribution-prevention managers to publicly disclose annual compliance reports (as described in Section 2.1). However, participants emphasized that despite this mandate, the reports lack clear criteria for how platforms evaluate takedown requests, categorize violations, or justify non-action. For example, many unresolved cases are labeled simply as "other," without explanation. Participants argued that beyond requiring disclosure, stricter and more focused standards are needed to ensure full accountability.

P4 further elaborated on the difficulties of contacting platforms hosted outside Korea: *"We send takedown requests in Chinese for sites hosted in China, and in English for others, but outside of Korean platforms, we rarely get a response."* Participants reported that Instagram, in particular, makes delayed replies—sometimes taking one to two months—and only responding in cases involving a child (P4). This delay severely undermines timely interventions, especially when platforms are hosted overseas or operate via VPN-masked servers.

Participants expressed that such poor cooperation from platforms undermines not only investigation efforts but also police motivation (P2, P3, P4, P7, P8, P9, P10). P4 stated,

> *The police often tell victims from the very beginning, 'We probably won't be able to catch the perpetrator,' because they know platforms won't respond. This makes it nearly impossible for survivors to even begin the recovery process. (P4)*

As a result, perpetrators grow increasingly confident that they will not be held accountable. This resignation leads to systemic inaction, emboldening perpetrators who believe they will not be caught (P8, P10).

These findings point to the urgent need for cross-platform governance mechanisms that standardize reporting, remove the burden of verification from survivors, and ensure transparent and timely platform responses.

4.2.3 ***Design Suggestion 2: Creating a Case File Rather Than Reporting an Incident.*** Most current reporting systems on social media platforms treat each abusive item (photo or video) as a standalone case, requiring users to submit individual takedown requests per content item. However, participants emphasized that what is needed during investigations is not merely the removal of isolated content, but documentation of the entire dissemination context. As P7 noted, it is *"the record of how the content was shared and who was involved*" that enables legal accountability. Therefore, an integrated case management system was proposed—one that consolidates all abuse-related activities into a unified case file rather than isolated tickets.

This approach is particularly critical in the context of deepfake sexual abuse, where perpetrators can deny responsibility by claiming resemblance is coincidental. As P10 emphasized, *"Chat logs and sharing behavior are the only way to prove intent, so gathering that evidence in one place is essential."* Such a system would substantially reduce the burden currently placed on survivors, activists, and legal professionals who must rely on manual or undercover investigations to gather evidence. Perpetrators often delete chat logs, accounts, or shared content once investigations begin, making intent difficult to prove. Undercover monitoring—sometimes required for months or even years [60]—demands immersion in exploitative environments, delaying justice and exposing investigators to significant psychological harm, including vicarious trauma and PTSD. However, implementing a centralized case file system raises ethical and technical concerns, particularly regarding data access and user privacy. Accessing deleted content or chat logs could potentially lead to over-surveillance or misuse of sensitive data. For example, if such systems are not carefully regulated, there is a risk that governments or other actors could request personal data under unrelated pretexts. To mitigate this, any cross-platform case management system should be restricted to specific abuse contexts, such as CSAM, NCII, or deepfake sexual abuse, and governed by strict privacy safeguards and due process mechanisms.

In all workshop sessions, participants proposed a linked reporting infrastructure where a single submission would trigger content checks across multiple platforms. This idea was grounded in the reality of cross-platform abuse: perpetrators collect a victim's photo from Instagram, request deepfake creation via Threads or X, and disseminate the resulting content on Telegram. P4 explained, *"Just one image can be used to generate multiple deepfakes that end up across various platforms. Currently, we have to manually search them one by one. If we could search by putting original image and search result will show across all platforms, we could collect evidences much faster."*

Participants also envisioned platform coordination in managing derivative content. P1 and P4 noted that even when the original content is deleted, platforms vary in how they treat reposts, screenshots, or links to the original post. An effective system would support cross-platform detection of variations and maintain a unified takedown process that recognizes these derivative forms (P4, P6, P7, P8, P9, P10).

In this suggestion, existing collaboration models offer insight into implementation pathways. Project Lantern [87] facilitates cross-platform sharing of behavioral signals for child sexual exploitation, while StopNCII.org [83] enables victims to hash intimate images for multi-platform removal. However, both rely on perceptual hashing, which performs poorly on altered images—e.g., cropped, filtered, or watermarked variants. This limitation is magnified in deepfake sexual abuse, where a single prompt can generate hundreds of visually distinct outputs that evade static hash-matching. Addressing this requires detection systems that go beyond hashing, incorporating adaptive ML-based similarity matching and contextual metadata analysis to recognize semantically equivalent content despite surface-level variation.

Also, there is remaining limitations on current collaboration models that they depend on voluntary platform participation, limiting their reach when companies do not actively engage. Policy frameworks such as the *EU Digital Services Act (DSA)* offer mechanisms for incentivizing compliance by requiring large platforms to assess systemic risks, mitigate illegal content, and cooperate with trusted flaggers, with non-compliance subject to substantial penalties [23]. Aligning case-based reporting infrastructure with such regulatory frameworks—while ensuring metadata standardization and third-party oversight—can foster more consistent platform participation and accountability, without undermining user privacy.

4.2.4 ***Design Suggestion 3: Platform Should Prove the Non-Existence of Deepfake Content.*** Participants expressed frustration with the current reporting process that places the burden of proof on survivors, requiring them to demonstrate the existence and harm of abusive content. Instead, **they called for a shift in responsibility: platforms should be required to proactively demonstrate that harmful content does not exist or has been fully removed.** This must reflect survivor-centered needs, specifically, enabling transparency, trackability, and assurances of safety within digital spaces. As P4 explained,

> *There are two goals in monitoring. First, to 'find all' sites where the content is being circulated for evidence collection; and second, 'find nothing' to confirm that it no longer exists anywhere, so the survivor can regain peace of mind. (P4)*

This confirmation plays a significant role in the survivor's psychological recovery. Even when technical deletion is imperfect—due to stored copies or reuploads—participants stressed the importance of establishing institutional trust that long-term monitoring will continue. P6 emphasized, *"Survivors need to see that even if abusive content reappears later, there are systems in place to catch and remove it again. That assurance is critical for recovery. (P6)"*

To support this, participants proposed several features: platforms should transparently report what actions were taken in response to a report, including how many items were found and removed across

platforms (P1, P3). Survivors should also receive regular updates on monitoring outcomes, and have the option to unsubscribe the report once they feel safe (P8). These efforts should not be positioned as optional customer service but as core responsibilities of platforms committed to user safety. To strengthen feasibility and shift power toward survivors, reporting systems should provide tools that reduce manual tracking and increase transparency. Building on StopNCII.org's company-side interface, a survivor-facing dashboard could display which platforms reviewed submitted hashes and what actions they took. Non-responses could trigger survivor follow-ups or be publicly flagged. Publishing platform-level response metrics—such as review or action rates—would support individual survivors while enabling civil society to monitor institutional responsiveness and pressure underperforming companies.

4.2.5 ***Design Suggestion 4: International Collaborative Investigation System.*** Given the global nature of online platforms and the borderless spread of deepfake content, participants argued for an international infrastructure dedicated to collaborative investigation and enforcement, scaling up earlier design suggestions into a global level. P6 and P8 highlighted that national laws vary significantly in how sexual abuse is defined and prosecuted, which makes uniform responses challenging. For example, under South Korean law, it can be difficult to criminalize deepfake sexual content if the depicted individual cannot be clearly identified using real images. In contrast, jurisdictions like Saskatchewan and British Columbia in Canada criminalize such content regardless of whether the depicted person is identifiable [54]. Similarly, in the U.S., a Virginia statute extends protections to individuals who merely resemble the victim in generated content[8]. These legal discrepancies highlight the complex and evolving landscape of deepfake regulation, and underscore the difficulty of coordinating timely and effective interventions across borders. To address these challenges, participants envisioned an AI-supported system capable of learning each country's legal definitions and social context to flag and process content accordingly (P8). This system could serve as an intermediary infrastructure to match reported content with jurisdiction-specific standards, helping platforms, NGOs, and law enforcement agencies determine the appropriate course of action based on legal viability. Moreover, it could incentivize platforms to proactively comply with local regulations, as the AI system can dynamically adapt to changes in legal frameworks. This would reduce the burden on platforms to manually update internal protocols or allocate additional legal and engineering resources each time new regulatory requirements are introduced. The technical challenges in developing such a system are significant; modern AI models have shown significant inaccuracies when applied to legal contexts [17]. However, participants' vision for a future AI-driven tool to help navigate international legal complexities highlights the need for better supporting infrastructure in this regard.

Extending this concept, participants called for the establishment of an independent international investigative body composed of technical experts, with the authority and resources to intervene across jurisdictions. As P2 argued, "This agency should not just coordinate but have real investigative powers and funding, so it can create a one-stop system for new forms of digital sexual violence." An international investigative infrastructure could also ease the burden on survivors and NGOs navigating fragmented legal systems. Deepfake sexual abuse and broader NCII content are frequently hosted on overseas illegal sites, making IP tracking slow or infeasible. Even when traced, Korean police must initiate lengthy cross-border cooperation during which perpetrators often erase digital evidence. These challenges highlight the need for a coordinated global system capable of tracing not only individual cases but the broader networks and profit structures behind deepfake sexual exploitation. Such governance would enable proactive investigation, cross-jurisdictional enforcement, and long-term accountability.

4.3 Supporting activists against secondary trauma.

4.3.1 ***Activists at Risk of Deepfake Threats.*** Participants emphasized that the nature of deepfake technology increasingly threatens the safety of those who support survivors. While previous forms of retaliation against activists were limited to malicious comments or threatening messages on social media, these have now escalated into direct attacks using deepfake content. As P5 explained, *"If an activist's identity is exposed, perpetrators create and share deepfake content using their image."* Similarly, P2, who had appeared publicly in the media, shared that she received deepfake videos featuring her face via private messages as a form of intimidation. Unlike past threats, deepfake sexual abuse now exposes not only survivors but also activists to direct victimization, placing them in the dual position of supporter and potential target.

4.3.2 ***Monitoring Demands Long-Term Immersion in Abusive Contexts.*** Participants (P2, P4, P7, P10) described the burdensome and emotionally taxing nature of monitoring deepfake sexual abuse content, which requires manually searching across multiple platforms and prolonged exposure to harmful environments. As P6 noted, even after content is taken down, perpetrators frequently reupload stored files, necessitating continuous, long-term monitoring. Due to the limited availability of public investigators to carry out this work, survivors, activists, and their acquaintances are often left to assume this responsibility. Some activists reported infiltrating Telegram chatrooms and building rapport with perpetrators to gather evidence.

This process demands adaptation to abusive online cultures, which participants described as deeply distressing. Telegram was frequently cited for its multi-layered structure: so-called "link hubs" share access to hidden groups, where new users must first "prove" themselves to gain entry to rooms where illegal content is circulated. P7 and P10 shared that they often had to agree with perpetrators' degrading comments toward victims to avoid raising suspicion. Others reported being coerced into uploading sexually explicit content (P7, P10), registering for gambling sites (P7), or submitting ID photos to verify they were male (P10). Even police investigators struggled to access these channels, often being asked school-specific questions that only real students could answer [69].

4.3.3 ***Exposure to Sexual Violence Content Causes Long-Term Psychological Harm.*** Across all workshops, participants identified long-term exposure to abusive content as a primary

[8]Virginia Code § 18.2-386.1, Unlawful creation of image of another; penalty.

source of secondary trauma and emotional exhaustion. Repeated encounters with sexual violence material—particularly in the context of monitoring and evidence collection—produced lasting psychological distress. As P6 shared, *"After starting this work, I began having recurring nightmares in which I was the victim. I've also developed difficulty forming social relationships with men."* Similarly, P2 noted, *"Some of the videos I watched three or four years ago still replay vividly in my mind."*

This trauma was exacerbated by the lack of platform accountability: content often reappears even after takedown requests, leading to a cycle of helplessness and isolation (P2, P3, P4, P5, P6, P7, P8, P10). Despite this, many activists remain in precarious, non-permanent employment positions (P3, P4, P5), with little institutional or psychological support. As a result, they face a profound mismatch between the emotional demands of their labor and the resources available to sustain it.

4.3.4 ***Design Suggestion 5: Supporting Activist Mental Health in Monitoring.*** To reduce the emotional burden of monitoring deepfake sexual abuse content, participants emphasized integrating AI to minimize direct exposure. They proposed systems that detect content using hash values or facial feature embeddings (e.g., eyes, nose, mouth) from victim images, enabling AI to pre-screen material, flagging potential matches for minimal human confirmation, thereby reducing trauma risk. (P2, P3, P4, P5, P6, P7, P10).

In addition to automation, participants emphasized the importance of building safe monitoring environments that protect activists' privacy. For instance, those engaged in evidence collection reported using VPNs and other anonymity tools to avoid location tracking by perpetrators (P4, P6, P10). To limit exposure, some activists intentionally stop viewing after 10 seconds once victim identification is confirmed (P2, P3, P4), while others structure their day to avoid extended periods of content review (P7, P10). Based on these practices, participants suggested interface features such as displaying content in grayscale by default to reduce visual intensity, timed monitoring alerts that prompt regular breaks, and mental health prompts embedded in the system to encourage reflection and rest. Rather than fully automating all decisions, these features aim to reduce harm by balancing efficiency with care.

Prior work has examined interface-level interventions—such as greyscaling and blurring—for platform-employed moderators at Google, Reddit, and other commercial platforms [18, 47]. Steiger et al. [82] further highlight the long-term psychological toll content moderation can take, advocating for moderator-centered systems of care. However, these discussions have yet to extend meaningfully to the NGO sector, where grassroots activists monitoring abusive content without institutional support, mental health resources, or technical infrastructure. To support NGO adoption of safe moderation tools, feasibility constraints must be addressed. Unlike profit platforms, NGOs lack engineering resources, stable funding, and access to proprietary moderation systems. Developing open-source, modular toolkits tailored to NGO workflows-e.g., grayscale viewing, exposure limits, and embedded mental health prompts-could provide accessible alternatives suited to low-resource environments. Cross-sector collaboration with research institutions and civic tech groups could further support co-development and capacity building. Policy support may also help, such as integrating activist protection criteria into digital-safety funding or mandating moderator wellbeing standards in government-supported monitoring programs.

5 Discussion: Toward Proactive and Responsible Ecosystems

Building on participants' design suggestions, this section explores key sociotechnical and ethical considerations that must be debated in the design of gernerative AI and social media platform moderation. We acknowledge that differences in privacy considerations—particularly between consensual sexual content and CSAM, as well as in platform data-sharing policies—make it difficult to offer concrete, immediately actionable design suggestions. Instead, zooming out from focusing on the technical intervention, we aim to expand the discussion to examine virtual harm, data ownership, and the role of activist communities in co-governing digital safety infrastructures.

5.1 Virtual but Violent?: Rethinking Harm in the Age of AI-Generated Sexuality

One of the central debates surrounding deepfake sexual abuse concerns whether harm exists in the absence of physical violence or real individuals. From a legal standpoint, possession of AI-generated abusive material sparked controversy over whether, if a person just generated the content without sharing, this could still be defined as a crime [19, 42]. Despite such debates, there is a growing parallel discourse that deepfakes, by not involving real women, make virtual pornography "safer." This argument is often tied to the emergence of AI-generated virtual characters used in dating apps, chatbots, and other intimate interactions, where no real individuals are directly harm [10, 55].

While recent HCI researchers has explored how AI-generated companions and avatars can reshape intimacy and dating practices [10, 55], Fan et el. shows that users perceive these systems as replicating discriminatory or stereotypical behaviors toward women. Broader reviews also raise concerns that gendered biases embedded in AI systems risk reinforcing structural inequalities[20, 36]. Legal scholars such as McGlynn et al. emphasize that central issue is not just a act of content creation, but the issue of power and control and humiliation of women. In this context, the greater risk lies not in over-criminalization, but in the continued under-criminalization of gendered harms that are normalized in digital spaces [61]. These concerns parallel our participants' fears that deepfake content—even when using synthetic or fictional characters—contributes to a culture that normalizes sexual objectification and dehumanization. For instance, P5 described how pornographic deepfakes of 3D female game characters are used as social currency in gaming communities: *"They trade them to get access to cheat codes or game tips."* While some argue these depictions are harmless since they involve virtual characters, they nonetheless reinforce harmful norms and sexual objectification [20, 24].

The increasingly realistic nature of generative AI enables what participants described as "zero-resistance" sexual interactions. Unlike real-world relationships that require consent, communication, and negotiation, AI-generated intimacy—whether through chatbots or synthetic pornography—offers users complete control. This can distort expectations about real relationships [20, 27, 36]. P4, P6, and

P8 observed that young perpetrators often start with deepfakes of fictional characters, then celebrities, and eventually real peers, illustrating a slippery slope toward targeting real individuals.

As P9 noted regarding a prominent university chatroom case: *"Their goal wasn't money or blackmail. They just wanted to humiliate women they knew. It was their form of entertainment."* Participants stressed that viewing deepfakes of acquaintances or celebrities serves not only as entertainment, but also as a means of degrading women's autonomy and status. The low barrier to creation and sharing facilitates a culture where gendered violence is trivialized as a game.

Participants also highlighted how deepfake content acts as both a gateway and incentive for further criminal behavior. As P2, P7, and P10 described, celebrity deepfakes were used as bait to lure users into Telegram channels, which then granted access to illegal footage of real women as a "reward" for activity. In these cases, deepfakes did not merely precede physical crimes; they were embedded in their ecosystem. This challenges the notion of a clear boundary between "virtual" and "real" abuse.

Grounded in the current ecosystem of deepfake sexual abuse, participants emphasized that societal perceptions of deepfakes as "not real" significantly undermine institutional and public responses. P1, who supports survivors within an institutional setting, described the disconnect: *"No matter how thoroughly I explained the harm that survivors are experiencing, the administrators just couldn't grasp it. Their passive responses made me feel powerless."* This lack of understanding extended to law enforcement. Multiple participants (P2, P4, P5, P6, P9, P10) reported that police reactions varied widely depending on officers' individual gender sensitivity and understanding of AI, leading to inconsistent outcomes.

Rather than focusing solely on who is depicted, harm assessment must account for how AI-generated content reinforces systemic gendered violence. The issue lies not only in consent or realism, but in how such materials normalize power asymmetries, sexual violence, and misogyny in digital cultures. Recent HCI work shows that public awareness of deepfake sexual abuse remains limited despite its growing prevalence [94], while AI-generated narratives can reflect and reinforce gendered struggle tropes that shape users' sense of identity and social belonging [25]. These dynamics point to deeper sociotechnical risks, where AI systems do not merely replicate data but actively shape and reinforce misogyny and gender hierarchy. Future research should investigate the psychological and social consequences of exposure to AI-generated sexual content. In particular, understanding how repeated interaction with synthetic intimacy shapes beliefs about consent, relationships, and gender is critical for informing both policy and design.

5.2 Who Owns Our Digital Selves?: Debates on Data Ownership and Post-Abuse Identity

Deepfake sexual abuse reveals an urgent gap in how personal data is governed in digital spaces. Participants repeatedly noted that once content is uploaded—whether by themselves or others—they lose meaningful control over it. This concern extends to AI-generated content, where one's image can be endlessly reproduced and weaponized without consent. Platforms that profit from user data rarely intervene, citing freedom of expression. As P9 put it, "Using the justification of protecting freedom of expression to allow nearly all behavior on platforms ultimately permits criminal activity as well. If platforms profit from user data, they must bear greater responsibility for the harms that arise within their ecosystems."

This loss of control deeply affects survivors' sense of self. Survivors attempt to reclaim autonomy offline through several actions, "including doing plastic surgery, changing their names, or applying for new national identification numbers. They live in constant fear that someone might recognize them from the circulated content. This creates identity confusion between who they were before and who they must become to avoid recognition (P2)." Even when survivors delete their accounts or step away from social platforms, traces of harm persist in screenshots, reposts, or archived data on perpetrators' devices. This signals a broader shift in how digital identity is governed—one that often places control in the hands of platforms rather than individuals.

This signals a broader shift in how digital identity is governed, one that often places control in the hands of platforms rather than individuals [52]. Brigham et al. [14] argue that traditional privacy frameworks, which focus on *data privacy*, are insufficient to address the harms of deepfake sexual abuse. They propose a shift toward *representational privacy* that accounts for generative AI generating sexual representations of individuals using minimal personal data. In this context, sexual consent must extend beyond personal data privacy to include control over one's actual image and representation. Building on this, Zytko et al. [102] explored participatory design of consent technologies with women and LGBTQ+ stakeholders in the context of online dating, and their work highlighted the need for consent to be understood as an ongoing, dialogic process rather than a one-time agreement. However, this approach remains centered on the notion of agreement, and has yet to fully address how structural power imbalances and platform-driven data extraction practices undermine users' ability to exercise meaningful agency.

In response to these concerns, participants pointed to emerging international efforts to rebalance control. In Denmark, lawmakers are discussing legislation that would grant individuals intellectual property rights over their faces and voices [3][9]. Such laws aim to prevent both unauthorized deepfakes and the non-consensual training of AI systems on personal data. Participants across all workshops expressed strong support for adopting similar protections in South Korea. They emphasized that personal data uploaded through social platforms should be recognized not as the property of those platforms but as the intellectual and human property of the individuals, regardless of whether it is used for training AI or even transformed through use of AI. This perspective reframes digital self as an extension of personal agency and calls for legal mechanisms that reflect that ownership.

At the technical level, detecting AI-generated content requires not only determining if manipulation occurred but also localizing

[9]The bill proposes amendments to the Danish Copyright Act: Section 65-a grants performing artists protection from unauthorized digital imitations of their performances; Section 73-a offers all individuals protection from unauthorized sharing of realistic deepfake representations of their face, voice, or body. If passed, it will allow citizens to claim copyright infringement over AI-generated depictions.

where—an essential step in protecting the data ownership of depicted individuals. Watermarking techniques such as Michael et al.'s Noise-Coded Illumination [62], which improve spatial and temporal watermarking, present promising directions for platform-level adoption. By enabling the localization of AI generated manipulations, such methods enhance the precision of forensic analysis. We highlight these advances to illustrate how AI platforms might incorporate stronger provenance signals as part of securing data ownership.

5.3 Lessons from Conducting Participatory Design Workshops with Activists

Civil society actors—including activists, victim support groups, and independent moderators—hold irreplaceable, ground-level knowledge of how harm unfolds in digital spaces. In our workshops, participants consistently demonstrated foresight in identifying not just present abuse patterns but emerging threats linked to evolving technologies. P3 noted, *"We started seeing deepfakes made from game characters as early as 2020, long before the media paid attention. It was already clear to us that generative AI would soon be used to target real women, and we raised alarms. But, as always, no one really listens until actual victims emerge and harm becomes undeniable."* As AI-generated sexual abuse continues to outpace formal regulatory responses, these actors function as early-warning systems, capable of predicting misuse and proposing proactive interventions before harm escalates.

Designing safeguards should therefore move beyond expert-driven or top-down policy development. Prior research in participatory design for safety technologies emphasizes the importance of survivor-centered co-design [50, 84]. D'Ignazio et al. suggests to involve activists in co-designing technology because activists' hidden work of moderating content, providing emotional support, organizing schedules, and other crucial tasks are essential and require deep expertise, but are not recognized or compensated [50]. Our findings revealed the extent of this invisible labor and the psychological burden involved in activists' monitoring. As shown in Section 4.3, activists endure emotional trauma, legal uncertainty, and high-stakes moderation work, often without institutional support. Yet, their insights into content sharing, perpetrator behavior, and platform gaps remain undervalued. Trauma-informed participatory design [5, 59] must recognize these roles—not only to improve tooling, but to reveal their underestimated labors and to capture their challenges from labor being invisible.

To meaningfully embed marginalized actors in the governance of content moderation, institutional stakeholders must commit to sustained structural support. Our participatory workshops revealed that effective co-design depends on facilitators and researchers developing deep familiarity with the day-to-day workflows, constraints, and emotional labor of activist communities. This process cannot be reduced to one-off consultations. We call for recurring participatory audits, transparent reporting collaborations, and formal inclusion of survivor and activist voices in the strategic planning of platform safety efforts. These actors must be recognized not as temporary advisors, but as co-governors of the digital infrastructures they help to monitor, repair, and reimagine.

6 Limitations

We acknowledge several limitations of the current study. First, all participants were based in South Korea, which may have shaped their perspectives through specific cultural, legal, and institutional lenses. Consequently, the identified challenges and design suggestions may not generalize to regions with different infrastructures. For instance, design suggestions related to case reporting systems (section 4.2.3) reflect South Korean legal contexts and may not translate directly to other regions. With this said, research and journalism have consistently shown that deepfake sexual abuse is a truly global problem—not a problem limited to South Korea. Future cross-cultural research is needed to examine how diverse legal and technological environments influence responses differently to deepfake sexual abuse.

Second, we were unable to recruit police investigators as participants due to access constraints. While legal professionals may follow different procedures, our study intentionally foregrounded activists' perspectives—guided by D'Ignazio et al. [84]—who step in the gaps left by incomplete, inaccessible, or under-resourced governmental systems. These activists play a critical role in monitoring, identifying victims, and shaping responses in under-reported and stigmatized contexts.

Lastly, the *speculative design* activity (section 3.4.4) included predefined stakeholder and technology cards, which may have limited participant creativity. To address this, facilitators explicitly encouraged participants to imagine interventions without concern for current technical feasibility, urging them to assume "any technology is possible." However, this in turn may have affected the direct technological feasibility of proposed interventions. For this reason, we emphasize that the design suggestions noted in the results section should be interpreted as guiding objectives rather than specific technical proposals.

7 Conclusion

This research examined the distinct harms of deepfake sexual abuse and the challenges posed in current monitoring systems, and proposed survivor-centered infrastructure design directions. Through participatory design workshops with survivors and activists engaged in victim advocacy and digital monitoring, we co-designed proactive safeguards, a shift in platform responsibility to prove the absence of abuse, and coordinated, cross-platform approaches to evidence collection and takedown. We also underscore the urgent need for structural support and mental health protections for activists who monitor for this type of content, whose labor remains largely invisible and unsupported. These insights demonstrate the need for a broader societal conversation about how AI-generated sexual content should be addressed through organizational, legal, and technical means. We also show how this content intersects with and reinforces older forms of gendered violence and how it demands new definitions of bodily data ownership.

Acknowledgments

We thank facilitators Naeun Choi and Sarang Kwon for their feedback to the workshop design and facilitation. Their expertise in survivor-centered facilitation greatly strengthened the study. We are deeply grateful to all study participants, as well as the activists

who generously shared their experiences, connected us with organizations, and provided critical insights throughout the research. We also thank the members of the Collaborative Social Technologies Lab (CSTL) at KAIST for their encouragement, thoughtful feedback, and continued support throughout this project.

We especially express our respect and admiration for the activists in South Korea who continue to make digital sexual abuse visible, support victims, and advocate for justice in the face of ongoing harm.

References

[1] 2022. *Cambridge-based Internet Watch Foundation warns of rise in remote abuse.* https://www.bbc.com/news/uk-england-cambridgeshire-61191906

[2] 2023. *IWF warning over use of AI-generated abuse images.* https://www.bbc.com/news/uk-england-cambridgeshire-67145583

[3] 2025. *Denmark's Deepfake Legislation: Bold Copyright and Digital Identity Protection.* https://abounaja.com/blog/denmarks-deepfake-legislation-bold-copyright-and-digital-identity-protection

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–7. doi:10.1109/WIFS.2018.8630761 ISSN: 2157-4774.

[5] Naseem Ahmadpour, Lian Loke, Carl Gray, Yidan Cao, Chloe Macdonald, and Rebecca Hart. 2023. Understanding how technology can support social-emotional learning of children: a dyadic trauma-informed participatory design with proxies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. doi:10.1145/3544548.3581032

[6] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace* 27 (2019). https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[7] Mahrus Ali, Zico Junius Fernando, Chairul Huda, and Mahmutarom Mahmutarom. 2025. Deepfakes and Victimology: Exploring the Impact of Digital Manipulation on Victims. *Substantive Justice International Journal of Law* 8, 1 (May 2025), 1–12. doi:10.56087/substantivejustice.v8i1.306

[8] Amnesty International. 2023. *South Korea: Google Fails to Tackle Online Sexual Abuse Content Despite Complaints by Survivors.* https://www.amnesty.org/en/latest/news/2023/12/south-korea-google-fails-to-tackle-online-sexual-abuse-content-despite-complaints-by-survivors/ Accessed September 12, 2025.

[9] AMPOS. 2025. *Material Seminar Detail.* https://ampos.nanet.go.kr/materialSeminarDetail.do?control_no=PAMP10000000076217 AMPOS seminar listing.

[10] Dünya Baradari, Tejaswi Polimetla, and Pattie Maes. 2025. Data-Driven AI Avatars for Valuation in Dating Scenarios. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3706599.3719833

[11] BBC News Korea. 2024. *Telegram Deepfake Sex Crimes: Why Were So Many Perpetrators Teenagers? (translated).* https://www.bbc.com/korean/articles/c4gl8zpd7e0o Accessed August 2024.

[12] Rosanna Bellini, Emily Tseng, Noel Warford, Alaa Daffalla, Tara Matthews, Sunny Consolvo, Jill Woelfer, Patrick Kelley, Michelle Mazurek, Dana Cuomo, Nicola Dell, and Thomas Ristenpart. 2024. SoK: Safer Digital-Safety Research Involving At-Risk Users. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP '24)*. 635–654. doi:10.1109/SP54263.2024.00071

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* (Jan. 2006). https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa Publisher: Taylor & Francis Group.

[14] Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. 2024. "Violation of my body": Perceptions of AI-Generated Non-Consensual (Intimate) Imagery. In *Proceedings of the Twentieth USENIX Conference on Usable Privacy and Security (SOUPS '24)*. USENIX Association, USA, 373–392.

[15] Matt Burgess. 2023. *Millions of People Are Using Abusive AI 'Nudify' Bots on Telegram.* https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/ Section: tags.

[16] Eung-Hyeok Chang. 2024. Review of Introduction of Undercover Investigation into Deepfake Sex Crimes Targeting Adults. *Criminal Investigation Studies* 10, 3 (2024), 5–24. doi:10.46225/CIS.2024.12.10.3.5 Publisher: Criminal Investigation Institute at Korean National Police University.

[17] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts Towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2454–2469. doi:10.1145/3630106.3659048

[18] Brandon Dang, Martin J Riedl, and Matthew Lease. 2018. But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. *arXiv preprint arXiv:1804.10999* (2018).

[19] Rebecca Delfino. 2019. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review* 88, 3 (Dec. 2019), 887. https://ir.lawnet.fordham.edu/flr/vol88/iss3/2

[20] Wen Duan, Lingyuan Li, Guo Freeman, and Nathan McNeese. 2025. A Scoping Review of Gender Stereotypes in Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–20. doi:10.1145/3706598.3713093

[21] Anthony Dunne and Fiona Raby. 2013. *Speculative Everything: Design, Fiction, and Social Dreaming*. The MIT Press.

[22] Pinterest Engineering. 2021. *Fighting spam with Guardian, a real-time analytics and rules engine.* https://medium.com/pinterest-engineering/fighting-spam-with-guardian-a-real-time-analytics-and-rules-engine-938e7e61fa27

[23] European Commission. 2025. *Digital Services Act: keeping us safe online.* https://commission.europa.eu/news-and-media/news/digital-services-act-keeping-us-safe-online-2025-09-22_en Press release, European Union.

[24] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713477

[25] Aidan Z. Fitzsimons, Elizabeth M. Gerber, and Duri Long. 2025. AI constructs gendered struggle narratives: Implications for self-concept and systems design.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2290–2301. doi:10.1145/3715275.3732156

[26] Camille François, Juliet Shen, Yoel Roth, Samantha Lai, Mariel Povolny, In M L Daniel, A Menking, M T Savio, and J Claffey. 2025. Four Functional Quadrants for Trust & Safety Tools: Detection, Investigation, Review & Enforcement (DIRE). (July 2025).

[27] Guo Freeman, Kelsea Schulenberg, Lingyuan Li, Ruchi Panchanadikar, and Nathan McNeese. 2025. "Comforting and Small Like a House Cat, Big and Intimidating Like a Bodyguard": How Women Perceive and Envision AI Companions as a New Harassment Mitigation Approach in Social VR. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3706598.3713473

[28] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (Dec. 1996), 16–23. doi:10.1145/242485.242493

[29] Shelby Grossman, Rachel Pfefferkorn, David Thiel, Shoren Shah, Renée DiResta, Jack Perrino, Eliza Cryst, Alex Stamos, and Jeff Hancock. 2024. *The Strengths and Weaknesses of the Online Child Safety Ecosystem.* Technical Report. Stanford Digital Repository. doi:10.25740/pr592kc5483 Accessed September 12, 2025.

[30] Gabriella De Guzman. 2025. *AI-generated child sexual abuse: The new digital threat we must confront now.* https://www.thorn.org/blog/ai-generated-child-sexual-abuse-the-new-digital-threat-we-must-confront-now/

[31] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric. 2025. Characterizing the {MrDeepFakes} Sexual Deepfake Marketplace. In *34th USENIX Security Symposium (USENIX Security 25)*. 5169–5188. https://www.usenix.org/conference/usenixsecurity25/presentation/han

[32] Jeffrey T. Hancock and Jeremy N. Bailenson. 2021. The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (March 2021), 149–152. doi:10.1089/cyber.2021.29208.jth Publisher: Mary Ann Liebert, Inc., publishers.

[33] Hankook Ilbo. 2022. *Requested Google to Delete 'Sexual Exploitation Material'...It Took a Year Just to Get a Response (translated).* https://www.hankookilbo.com/News/Read/A2022120715180003262 Accessed September 12, 2025.

[34] Will Hawkins, Brent Mittelstadt, and Chris Russell. 2025. Deepfakes on Demand: The Rise of Accessible Non-Consensual Deepfake Image Generators. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1602–1614. doi:10.1145/3715275.3732107

[35] Melissa Heikkilä. 2025. *Bans on deepfakes take us only so far—here's what we really need.* https://www.technologyreview.com/2024/02/27/1089010/bans-on-deepfakes-take-us-only-so-far-heres-what-we-really-need/

[36] Sumin Heo, Erika R Chen, and Jasmine Khuu. 2025. Exploring Gender Biases in LLM-based Voice Chatbots for Job Interviews. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–8. doi:10.1145/3706599.3719281

[37] Home Security Heroes. 2023. *2023 State of Deepfakes: Realities, Threats, and Impact.* https://www.securityhero.io/state-of-deepfakes/ Technical report.

[38] Tharon Howard. 2014. Journey mapping: a brief overview. *Commun. Des. Q. Rev* 2, 3 (May 2014), 10–13. doi:10.1145/2644448.2644451

[39] Hankook Ilbo. 2024. *Pay 1 Dollar and in a Minute You Can Create Porn...It's Easy to Make a "Deepfake Bot" but Hard to Punish (translated).* https://www.hankookilbo.com/News/Read/A2024082912390004564 Accessed August 29, 2024.

[40] Hankook Ilbo. 2025. *1,807 victims of "deepfake sex crimes" identified by the National Center for Digital Sexual Crime Response (NCDSCR) over the past year...a 128% increase (translated).* https://www.hankookilbo.com/News/Read/A2025082810070004233 News article, Section: Society.

[41] Soyoun Jang, Jay David Bolter, Richmond Y. Wong, Heidi Biggs, Robert Soden, Vera Khovanskaya, Laura Forlano, and Sasha de Koninck. 2025. Expanding Historical Approaches to Speculative Design. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference (DIS '25 Companion)*. Association for Computing Machinery, New York, NY, USA, 26–29. doi:10.1145/3715668.3734164

[42] Hyun Kyong Joo. 2024. Social Risks of Deepfakes and Their Regulatory Approaches: Focusing on Deepfake-related Crimes. *Gachon Law Review* 17, 3 (2024), 261–306. doi:10.15335/GLR.2024.17.3.008 Publisher: Law Research Institute.

[43] Journal of Online Trust and Safety. 2021. An Overview of Perceptual Hashing. *Journal of Online Trust and Safety* (2021). https://tsjournal.org/index.php/jots/article/view/24 Accessed September 12, 2025.

[44] Cecilia Kang. 2025. A.I.-Generated Images of Child Sexual Abuse Are Flooding the Internet. *The New York Times* (July 2025). https://www.nytimes.com/2025/07/10/technology/ai-csam-child-sexual-abuse.html

[45] Dr. Vasileios Karagiannopoulos, Dr. Annie Kirby, Shakiba Oftadeh-Moghadam, and Dr. Lisa Sugiura. 2021. Cybercrime awareness and victimisation in individuals over 60 years: A Portsmouth case study. *Computer Law & Security Review* 43 (Nov. 2021), 105615. doi:10.1016/j.clsr.2021.105615

[46] Clare-Marie Karat, John Karat, and Carolyn Brodie. 2005. Editorial: why HCI research in privacy and security is critical now. *Int. J. Hum.-Comput. Stud.* 63, 1-2 (July 2005), 1–4.

[47] Sowmya Karunakaran and Rashmi Ramakrishnan. 2019. Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.

[48] Songee Kim and Wonjin Kim. 2025. *Submitting a 'Deepfake Prevention Pledge' Before Taking Graduation Photos… The Ministry of Education Distributes a Digital Sex Violence 'SOS Guide' (translated).* https://www.khan.co.kr/article/202504221200001 Section: Society.

[49] Thomas C. King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. 2020. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics* 26, 1 (Feb. 2020), 89–120. doi:10.1007/s11948-018-00081-0

[50] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness Accountability and Transparency*. ACM, Rio de Janeiro Brazil, 100–112. doi:10.1145/3630106.3658543

[51] Pavel Korshunov and Sebastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. doi:10.48550/arXiv.1812.08685 arXiv:1812.08685 [cs].

[52] Kalle Kusk and Midas Nouwens. 2025. How Website Owners Use Consent Management Platforms: An Interview Study. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3706599.3720002

[53] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. 2021. Bad machines corrupt good morals. *Nature Human Behaviour* 5, 6 (June 2021), 679–685. doi:10.1038/s41562-021-01128-2 Publisher: Nature Publishing Group.

[54] Langlois Law LLP. 2024. *Legal Framework for Artificial Intelligence: What Are the Statutory Protections Against Deepfakes?* https://langlois.ca/en/insights/legal-framework-for-artificial-intelligence-what-are-the-statutory-protections-against-deepfakes/ Accessed September 12, 2025.

[55] Mateo Larrea, Xingyi Zhang, and Xuyang Zhu. 2025. LoveSims: Exploring 'What-If' Scenarios for Relationship Insights and Compatibility. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–7. doi:10.1145/3706599.3720011

[56] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. 3207–3216. https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html

[57] Kweilin T. Lucas. 2022. Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology. *Victims & Offenders* 17, 5 (July 2022), 647–659. doi:10.1080/15564886.2022.2036656

[58] Helen Burchell Lydia Dowling Ranera. 2025. AI puts real child sex victims at risk, IWF experts say. (June 2025). https://www.bbc.com/news/articles/cgeqdxyqvexo

[59] Richard Martinez, Mark Van Hollebeke, Kurt Squire, Tong Wu, and Marco Zamarato. 2025. Generative Dreams: Rethinking GenAI Design Through a Community-based Approach with Recently Incarcerated, Gang Affiliated, and At-risk Young Adults at a Trauma Informed Arts Center. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3706599.3706679

[60] MBC. 2024. *Exclusive Report: "Disappointed by Powerless Police" — Victim Who Became Part of the Tracking Collective 'Bulgkot' (translated).* https://imnews.imbc.com/replay/2024/nwdesk/article/6628395_36515.html Accessed September 12, 2025.

[61] Clare M. S. McGlynn and Rüya Tuna Toparlak. 2024. The New Voyeurism: Criminalising the Creation of "Deepfake Porn". (July 2024). https://ssrn.com/abstract=4894256 Accepted for publication in the Journal of Law and Society.

[62] Peter Michael, Zekun Hao, Serge Belongie, and Abe Davis. 2025. Noise-Coded Illumination for Forensic and Photometric Video Analysis. *ACM Trans. Graph.* 44, 5, Article 165 (June 2025), 16 pages. doi:10.1145/3742892

[63] Microsoft. 2025. *PhotoDNA.* https://www.microsoft.com/en-us/photodna Accessed September 12, 2025.

[64] Midjourney. 2025. *Community Guidelines.* https://docs.midjourney.com/hc/en-us/articles/32013696484109-Community-Guidelines Midjourney documentation: community guidelines.

[65] Minyoung Moon. 2024. Digital Sex Crime, Online Misogyny, and Digital Feminism in South Korea. *Georgetown Journal of International Affairs* 25, 1 (2024), 186–192. https://muse.jhu.edu/pub/1/article/934902 Publisher: Johns Hopkins University Press.

[66] Lydia Morrish. 2024. *GitHub's Deepfake Porn Crackdown Still Isn't Working.* https://www.wired.com/story/githubs-deepfake-porn-crackdown-still-isnt-working/ Section: tags.

[67] National Center for Digital Sexual Crime Response (NCDSCR). 2025. *National Center for Digital Sexual Crime Response (NCDSCR).* https://d4u.stop.or.kr/ Accessed September 12, 2025.

[68] National Center for Missing & Exploited Children. 2025. *CyberTipline Data.* http://www.missingkids.org/content/ncmec/en/gethelpnow/cybertipline/cybertiplinedata.html Accessed September 12, 2025.

[69] NEWSTAPA. 2024. *The Accomplices of Deepfake Part 1: Infiltrating Telegram's "Layered Province" (translated).* http://newstapa.org/article/_NwuP Accessed September 12, 2025.

[70] Gabriel Tuhafeni Nhinda and Fungai Bhunu Shava. 2021. Towards the use of Participatory Methods in Cybersecurity research in rural Africa: A grassroots Approach. In *2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. 1–7. doi:10.1109/IMITEC52926.2021.9714649

[71] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. 7184–7193. https://openaccess.thecvf.com/content_ICCV_2019/html/Nirkin_FSGAN_Subject_Agnostic_Face_Swapping_and_Reenactment_ICCV_2019_paper.html

[72] OpenAI. 2025. *Creating images and videos in line with our policies.* https://openai.com/policies/creating-sora-videos-in-line-with-our-policies/ OpenAI policy page for Sora video and image generation.

[73] Pornhub. 2025. *Community Guidelines.* https://help.pornhub.com/hc/en-us/articles/4419900587155-Community-Guidelines Pornhub official community guidelines.

[74] Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa M. Redmiles. 2024. "Did They F***ing Consent to That?": Safer Digital Intimacy via Proactive Protection Against Image-Based Sexual Abuse. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA, 55–72. https://www.usenix.org/conference/usenixsecurity24/presentation/qin

[75] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642810

[76] Safer.io. 2022. *Tools to Detect CSAM and Child Exploitation.* https://safer.io/solutions/ Accessed September 12, 2025. Safer by Thorn.

[77] Kavous Salehzadeh Niksirat, Evanne Anthoine-Milhomme, Samuel Randin, Kévin Huguenin, and Mauro Cherubini. 2021. "I thought you were okay": Participatory Design with Young Adults to Fight Multiparty Privacy Conflicts in Online Social Networks. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 104–124. doi:10.1145/3461778.3462040

[78] Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices.* Lawrence Erlbaum Associates. Also available in paperback: ISBN 080580952X.

[79] Phoebe Sengers and Bill Gaver. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems*. ACM, University Park PA USA, 99–108. doi:10.1145/1142405.1142422

[80] Danny Slater and Betsy Masiello. 2023. Annex 2: Building Open Trust and Safety Tools. In *Scaling Trust on the Web: Comprehensive Report of the Task Force for a Trustworthy Future Web*. Atlantic Council. https://www.atlanticcouncil.org/in-depth-research-reports/report/scaling-trust/ Accessed September 11, 2025.

[81] Clay Spinuzzi. 2005. The Methodology of Participatory Design. *Technical Communication* 52, 2 (2005), 163–174.

[82] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators:

The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. doi:10.1145/3411764.3445092

[83] StopNCII.org. 2025. *StopNCII.org — Stop Non-Consensual Intimate Image Abuse.* https://stopncii.org/ Global tool to prevent sharing of non-consensual intimate images.

[84] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness Accountability and Transparency*. ACM, 667–678. doi:10.1145/3531146.3533132

[85] Theresa Jean Tanenbaum. 2014. Design fictional interactions: why HCI should care about stories. *Interactions* 21, 5 (Sept. 2014), 22–23. doi:10.1145/2648414

[86] Li Tang, Qingqing Ye, Haibo Hu, Qiao Xue, Yaxin Xiao, and Jin Li. 2024. DeepMark: A Scalable and Robust Framework for DeepFake Video Detection. *ACM Transactions on Privacy and Security* 27, 1 (Feb. 2024), 1–26. doi:10.1145/3629976

[87] Tech Coalition. 2023. *Announcing Lantern: The First Child Safety Cross-Platform Signal Sharing Program.* https://technologycoalition.org/news/announcing-lantern/ Tech Coalition press release.

[88] Tech Coalition. 2025. *Tech Coalition: Fighting Child Sexual Abuse Online.* https://www.technologycoalition.org/ Accessed April 26, 2025.

[89] The Guardian. 2025. *The rise of deepfake pornography in schools: 'One girl was so horrified she vomited'.* https://www.theguardian.com/society/ng-interactive/2025/dec/02/the-rise-of-deepfake-pornography-in-schools Accessed September 12, 2025.

[90] Thorn. 2023. *Safety by Design for Generative AI: Preventing Child Sexual Abuse and Exploitation.* Technical Report. Thorn. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf Accessed: 2025-12-03.

[91] Thorn. 2025. *Deepfake Nudes and Young People.* https://www.thorn.org/research/library/deepfake-nudes-and-young-people/ Thorn research publication, accessed online.

[92] Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg, and Rachel Greenstadt. 2023. Studying the Online Deepfake Community. *Journal of Online Trust and Safety* 2, 1 (Sept. 2023). doi:10.54501/jots.v2i1.126

[93] Bill Toulas. 2022. *Google quietly bans deepfake training projects on Colab.* https://www.bleepingcomputer.com/news/google/google-quietly-bans-deepfake-training-projects-on-colab/

[94] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. 2024. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642382

[95] United Nations Office on Drugs and Crime. 2025. *Deepfake Non-Consensual Intimate Material in Indonesia.* https://www.unodc.org/roseap/uploads/documents/indonesia/Deepfake_non-consensual_intimate_material_in_Indonesia_9.2025_EN_WEB.pdf UNODC report, accessed online.

[96] Vice. 2018. *Pornhub Bans Deepfakes.* https://www.vice.com/en/article/pornhub-bans-deepfakes/ Vice article on Pornhub's policy change.

[97] Gaurav Yadav, Md Zafar Sadique, Dr Suneel Kumar, Rachit Sharma, Dr Mamta Sharma, Dr Rama Sharma, and Dr Toshi Rattan. 2025. Psychological Trauma and Legal Challenges of Deep fake Technology. *Sciences of Conservation and Archaeology* 37, 1 (May 2025), 143–150. https://sci-arch.org/index.php/wwbhen/article/view/185

[98] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8261–8265. doi:10.1109/ICASSP.2019.8683164 ISSN: 2379-190X.

[99] C. Zauner. 2010. Implementation and Benchmarking of Perceptual Image Hash Functions. https://www.semanticscholar.org/paper/Implementation-and-Benchmarking-of-Perceptual-Image-Zauner/635e1b5261ad1545aab7acde48efa267ae428fc3

[100] Yuxiang Zhai, Xiao Xue, Zekai Guo, Tongtong Jin, Yuting Diao, and Jihong Jeung. 2025. Hear Us, then Protect Us: Navigating Deepfake Scams and Safeguard Interventions with Older Adults through Participatory Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3714423

[101] Annuska Zolyomi and Jaime Snyder. 2024. An Emotion Translator: Speculative Design By Neurodiverse Dyads. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–18. doi:10.1145/3613904.3642210

[102] Douglas Zytko and Nicholas Furlo. 2023. Online Dating as Context to Design Sexual Consent Technology with Women and LGBTQ+ Stakeholders. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3580911

A Workshop Recruitment Form

This workshop recruits two types of participants: (1) survivors of deepfake sexual abuse and (2) activists who support victims. Depending on your participation type, the survey questions will differ.

Definition of Deepfake Sexual Abuse: According to Article 14-2 of the South Korean Act on Special Cases Concerning the Punishment of Sexual Crimes, deepfake sexual abuse refers to editing, synthesizing, or altering recordings, videos, or voice recordings of a person's face, body, or voice in a sexually explicit or humiliating way without the individual's consent, and distributing or profiting from such content.

- Option 1: Aged 19 or older and a survivor of deepfake sexual abuse
- Option 2: Activist supporting victims of deepfake sexual abuse
- Option 3: Both a survivor and an activist

- (1) Name or Nickname
- (2) Email Address

If you selected Option 2 or Option 3:

- (3) Please select your organizational affiliation:
  - Police / Lawyer / Sexual Violence Counseling Center / Central Digital Sexual Crime Support Center / Local Prevention and Response Centers (Seoul, Gyeonggi, Incheon, etc.) / Digital Sexual Crime Relief Center / Other
- (4) Please write your occupation (including your organization).
- (4)-1 Please select your role at the organization:
  - Investigator / Counselor / Content Monitor / Takedown Support / Lawyer / Other
- (5) Select all areas related to your experience in responding to deepfake sexual abuse (based on categories provided by the Central Digital Sexual Crime Support Center). If your experience does not fall under any listed option, please specify under "Other."
  - Psychological & Emotional Support / Investigation & Legal Support / Medical Support / Takedown – Private SNS Platform (e.g., Telegram) / Takedown – Public Platforms (e.g., Twitter) / Monitoring – Private SNS Platform (e.g., Telegram) / Monitoring – Public Platforms (e.g., Twitter) / Policy Development & Consultation
- (5)-1 How many victims have you supported in this area?
  - Fewer than 5 / 5-10 / 10-30 / Over 30
- (6) Do you have experience responding to other types of digital sexual crimes? (Yes / No)
- (6)-1 If yes, how many victims have you supported?
  - Fewer than 5 / 5-10 / 10-30 / Over 30

If you selected Option 1 or Option 3:

This section asks about experiences of deepfake sexual abuse. You may skip any question you are uncomfortable answering. Responses are only used for workshop planning and session assignment. All data is encrypted and only accessible to the research team. Your

answers will not be quoted directly; findings will only be reported in aggregate.

- (7) Have you experienced deepfake sexual abuse? (Yes / No)
- (7)-1 What types of deepfake abuse were involved? (Select all that apply)
  - Synthetic content / Synthetic content with derogatory captions / Sent to me / Distributed via chat platforms / Distributed along with personal info / Used for threats / Used for extortion / Used to demanding sexual acts
- (7)-2 What actions were taken? (Select all that apply)
  - Police report / Contacted support center / Requested content takedown / Contacted platform / Legal consultation / Told friends or family / Hired digital undertaker / No action
- (7)-3 What was the biggest challenge in recognizing the abuse? (Select all that apply)
  - Unsure if it was abuse / Unable to identify the platform / Did not know where to seek help / Difficulty reporting / Difficulty contacting support centers / Takedown process too complex / Legal challenges
- (7)-4 What was the biggest challenge in responding to the abuse? (Select all that apply)
  - Unfamiliarity with procedures / Anxiety about identifying perpetrator / Fear of social response / Repeated victimization / Secondary harm / Financial burden
- (7)-5 Are you willing to share these responses during the workshop?
  - Yes / Yes, but only at survey level / No

Common Questions

- (8) What do you think is the most urgent task in responding to deepfake sexual abuse? (Select all that apply)
  - Early detection systems / Improved takedown processes / Perpetrator tracking / Evidence collection / Legal reform / Social awareness / Psychological support systems
- (9) Which of the following technologies have you heard of or are familiar with? (Used to adjust workshop content.)
  - Open-source AI model governance / Transparent AI training data / Platform content moderation / Real-time content filtering / AI-generated content detection / Watermark or hash-based content tracking / Multi-platform reporting system / Distributed monitoring involving stakeholders / Secure identity protection for victims / Monitoring systems for reviewer trauma / Photo misuse alert systems

B Workshop Pre-Survey Questions

- **Anonymous Participation for Participant Protection**
  - (1)-1. Will you turn on your camera during the workshop? (Yes / No)
  - (1)-2. Will you disclose your affiliated organization during the workshop? (Yes / No)
  - (1)-3. Will you participate anonymously (using a nickname) during the workshop? (Yes / No)
  - (1)-4. If you chose to participate anonymously, please provide the nickname you will use during the workshop:
- **(2) Which stage of responding to deepfake sexual abuse would you most like to discuss during the workshop?** (Multiple selections allowed)
  Options are based on the one-stop support stages defined by the Seoul Digital Sexual Crime Support Center. If your experience does not match any listed option, please specify under "Other."
  - Prevention
  - Recognition & Reporting
  - Investigation & Evidence Collection
  - Takedown Request
  - Monitoring of Content Distribution
  - Legal Support
  - Psychological & Medical Support
- **(3) How familiar are you with each of the following response stages to deepfake sexual abuse?** (Rate 1–5)
  - Prevention
  - Recognition & Reporting
  - Investigation & Evidence Collection
  - Takedown Request
  - Monitoring of Content Distribution
  - Legal Support
  - Psychological & Medical Support
- **(4)** Have you ever experienced difficulty because you didn't know how to respond to a case of deepfake sexual abuse?
- **(5)** Have you ever experienced difficulty because, although you knew how to respond, the available technology or support systems were insufficient?
- **(6)** Have you ever experienced difficulty because, although you knew how to respond, cooperation with investigative agencies (e.g., the police) was lacking?
- **(7)** Have you ever experienced difficulty because, although you knew how to respond, cooperation with other stakeholders (e.g., platform providers) was lacking?
- **(8)** Have you ever used any technology (software systems) in the process of responding to deepfake sexual abuse or supporting victims?
  'Technology' refers to systems like software platforms used by your organization, the Police Illegal Content Tracker, or AI-powered takedown request systems.
  - (8)-1. If yes, what technology did you use?
  - (8)-2. What effects did you expect from using the technology?
  - (8)-3. What were the actual effects of using the technology? If there were no effects, what challenges did you face?
- **(9)** What topics do you expect or hope to explore in the workshop'?

C Value Sensitive Design Framework Tag

Value	Definition
Human Welfare	Refers to people's physical, material, and psychological well-being.
Ownership & Property	Refers to a right to possess an object(or information), use it, manage it, derive income from it, and bequeath it.
Privacy	Refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others.
Freedom from Bias	Refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias
Universal Usability	Refers to making all people successful users of information technology.
Trust	Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal.
Autonomy	Refers to people's ability to decide, plan, and act in ways that they believe will help them achieve their goals.
Informed Consent	Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension(for 'informed') and voluntariness, competence, and agreement(for 'consent').
Accountability	Refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution.
Courtesy	Refers to treating people with politeness and consideration.
Identity	Refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time.
Calmness	Refers to a peaceful and composed psychological state.
Environmental Sustainability	Refers to sustaining ecosystems such that they meet the needs of the present without compromising future generations.

Table 2: Descriptions of Value-Sensitive Design (VSD) Tags used to support workshop discussion. These values helped participants articulate and group challenges during collaborative activities.

D Stakeholder & Technology Cards

E Qualitative Codebook

Stakeholder	Description
Victim Support Organizations	Organizations that provide psychological, legal, and medical assistance to victims of sexual crimes. They support victims through counseling, legal aid, recovery programs, reporting assistance, and advocating for victims' rights and recovery.
Social Media Platform	Companies that operate social media, search engines, and video platforms. They are responsible for preventing the spread of illegal content, maintaining reporting and takedown systems, and establishing user protection policies.
AI Developers and Companies	Individuals or organizations that develop and provide AI technologies. They are responsible for building safeguards to prevent misuse, adhering to data ethics, and creating technical countermeasures.
AI Open-Source Platforms	Platforms that provide open access to AI models and code. They have a responsibility to restrict harmful use of AI, implement abuse prevention policies, and maintain monitoring system
Police and Investigative Authorities	National institutions responsible for investigating sexual crimes, identifying perpetrators, collecting evidence, and protecting victims. They are expected to enhance cybercrime response capabilities and conduct timely investigations.
Government	Holds overall responsibility for legislation, policy-making, budget allocation, national victim protection, and preventing recurrence. Also responsible for coordinating inter-agency collaboration.
Education Sector (Schools)	Plays a role in prevention and response through sexual crime prevention education, digital citizenship and media literacy training for students and staff, and establishing support systems in case of incidents.

Table 3: Stakeholder roles in the digital sexual abuse response ecosystem. It was used as stakeholder cards during the workshop - speculative design activity.

Technology Category	Description	Example
Technologies for AI model management		
Open-Source AI Model Governance	Systems that control and manage the use of open-source AI models used to create deepfakes. This includes restricting model access, limiting usage purposes, and filtering outputs to prevent misuse.	"We want to develop a system that blocks requests to generate sexual deepfakes of specific individuals using text-to-image/video models."
Explainable AI	Technologies that disclose what data the AI model was trained on and how outputs are generated. Helps users understand the basis of AI decisions.	"Users should be informed if the training dataset includes CSAM or scraped data from harmful sites."
AI-Generated Content Detection	Tools to detect deepfakes via pixel, voice pattern, or facial movement analysis.	"Please verify whether the uploaded video is a deepfake."
Watermarking / Hash-Based Tracking	Embeds invisible identifiers into digital content to verify authenticity later.	"Automatically embed digital signatures into all camera-captured photos."
Technologies for Platform Operators		
Content Moderation Systems	Detects and removes inappropriate content through reports, automation, and human review.	"Automatically block hate speech or abusive posts in our community."
Real-Time Harmful Content Filters	Blocks harmful content at the time of upload using AI systems.	"If inappropriate scenes appear in livestreams, immediately stop the broadcast."
Cross-Platform Reporting	Unified reporting system for multiple platforms.	"A Facebook report should trigger checks and removal on Instagram and YouTube."
Photo Misuse Alerts	Notifies users if their photo is reused or manipulated online.	"Alert me if someone captures and misuses my profile picture."
Technologies for Investigating Sexual Abuse Material		
Distributed Monitoring Systems	Collaboration among NGOs, platforms, and law enforcement for monitoring.	"Enable simultaneous reporting to civic groups, police, and platforms."
Privacy-Protecting Tech for Victims	Encryption tools for anonymous reporting and secure evidence submission.	"Allow me to report deepfake abuse anonymously and safely."
Trauma-Informed Interfaces	Designs that reduce psychological harm, like grayscale or text filters.	"Please grayscale CSAM media to reduce trauma for moderators."

Table 4: Technology solutions used to address digital sexual abuse, categorized by function. These cards were used during the workshop speculative design activity.

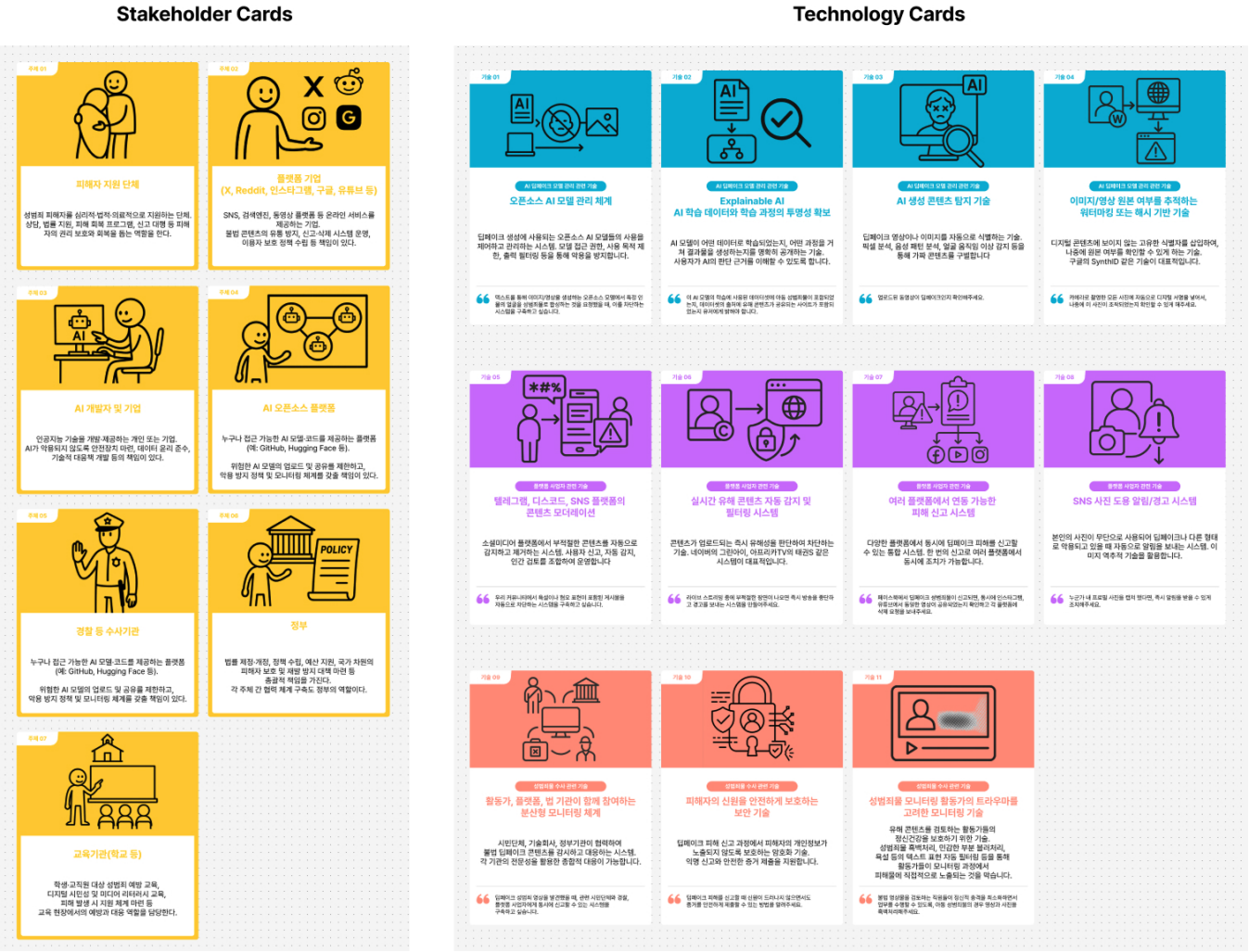


Figure 5: Stakeholder and Technology Cards used in the speculative design activity. Stakeholder cards (left, yellow) represent key actors in the ecosystem responding to deepfake sexual abuse. Technology cards (right, blue/purple/red) represent possible intervention technologies discussed during the workshop.

Category	Sub-Theme	Code	Description
Challenges	Ecosystem-Level Gaps	Platform Non-Cooperation	Online platforms' refusal or inability to provide access to evidence and cooperate with law enforcement investigations.
		Jurisdictional Limitations	Legal and technical barriers imposed by overseas servers and international boundaries limiting law enforcement capacity.
		Resource Scarcity	Insufficient government budget, resources, and institutional support for specialized deepfake response capacity.
		Institutional Siloing	Lack of coordination and information sharing between police, support services, educational institutions, and civil society organizations.
		Leadership Disconnect	Administrative and managerial failure to understand severity and urgency of digital sexual crimes.
	Technological Barriers	Evidence Access Obstacles	Difficulty collecting and documenting evidence due to encrypted platforms, paid access requirements, and anonymization tools.
		Perpetrator Anonymization	Use of VPN, anonymous accounts, and distributed networks making perpetrator identification nearly impossible.
		Source Code Availability	Open-source AI models and tools publicly available for creating deepfakes, distributed through GitHub and online forums.
		Slow Institutional Technology Adoption	Law enforcement agencies lacking technical capacity and training to work with emerging AI-based crime tools.
	Procedural Inefficiencies	Repetitive Victim Statements	Requirement for victims to repeat their accounts multiple times across different agencies and investigators.
		Delayed Case Processing	Prolonged investigation timelines and slow progression through legal procedures causing prolonged victim suffering.
		Inconsistent Legal Application	Variable enforcement of laws and sentencing disparities, particularly for minor perpetrators.
		Documentation Complexity	Complex procedures for official reporting and documentation creating barriers for victims.
	Psychosocial Harms	Secondary Victimization	Retraumatization of victims and monitoring activists through mandatory content viewing, repeated interviews, and public disclosure during evidence gathering.
		Loss of Human Trust	Victim experiences of betrayal when intimate relationships or peer connections are exploited to create deepfakes.
		Activist Burnout	Severe psychological fatigue experienced by counselors and activists from repeated exposure to harmful content and ineffective interventions.
		Powerlessness	Overwhelming sense of helplessness when institutional systems fail to deliver justice despite available evidence.
		Privacy Anxiety	Victim concerns about surveillance, data security, and ongoing exposure of personal information throughout investigation.
	Systemic Inequality	Gender Sensitivity Gaps	Law enforcement and institutional officials' lack of understanding about gender dynamics in sexual crimes and victim trauma.
		Societal Normalization	Cultural perception of deepfake creation as entertainment rather than crime, particularly among youth perpetrators.
		Inadequate Youth Accountability	Legal system's leniency toward minor perpetrators resulting in minimal consequences and continued victimization risk.
		Victim-Blaming Attitudes	Societal and institutional tendency to attribute responsibility to victims rather than perpetrators.
Speculative Design Ideas	Victim-Centered Goals	Proactive Crime Prevention	Technological and educational interventions designed to prevent deepfake creation before victimization occurs.
		Efficient Victim Reporting	Streamlined, one-time reporting systems enabling victims to provide evidence once to multiple agencies simultaneously.
		Rapid Content Removal	Automated systems enabling swift detection and deletion of deepfake material from all platforms within hours.
		Perpetrator Accountability	Technology and procedures ensuring consistent identification, prosecution, and appropriate sentencing of offenders.
		Comprehensive Victim Rehabilitation	Integrated trauma-informed support including psychological counseling, medical care, and social reintegration assistance.
	Multi-Stakeholder Framework	State Law Enforcement Leadership	Government police and prosecutors serving as coordinating authority with specialized deepfake investigation units.
		Platform Company Responsibility	Tech platforms implementing content moderation, data sharing, and perpetrator investigation cooperation.
		AI Developer Accountability	AI researchers and developers implementing ethical constraints and monitoring systems to prevent abuse.
		Educational Institution Integration	Schools implementing prevention curricula and institutional protocols for responding to student perpetrators.
		Civil Society Partnership	NGOs and victim support organizations providing specialized services and monitoring complementary to state capacity.
	Technological Mechanisms	AI-Based Content Filtering	Automated detection systems using facial recognition and pattern analysis to identify deepfakes across platforms.
		Victim Identification Systems	Image-based technology automatically aggregating all deepfake instances of same victim across platforms.
		Watermarking and Hash Tracking	Embedding unique identifiers in AI models and content enabling perpetrator attribution and content replication tracking.
		Source Code Detection	Technology identifying the specific AI model and source code used to create deepfakes for perpetrator linkage.
		Distributed Monitoring Network	Decentralized system of civil society and law enforcement monitoring with AI-powered analysis and integration.
		Integrated Case Management	Centralized digital system tracking victim information, evidence, and case status across all agencies.
	Policy & Governance	Ethical AI Development Standards	Mandatory requirements for AI developers including safety testing, dataset transparency, and abuse prevention mechanisms.
		Intellectual Property for Biometric Data	Legal framework assigning property rights to individuals' facial features and voice preventing unauthorized commercial use.
		Platform Compliance Requirements	Regulatory mandates requiring platforms to implement content filtering, data sharing, and investigation cooperation.
		Specialized Police Training	Mandatory education programs for law enforcement on deepfake technology, victim trauma, and gender-sensitive investigation.
		Prevention Education Campaigns	Community awareness programs teaching youth about legal consequences and ethical dimensions of deepfake creation.
		International Cooperation Frameworks	Cross-border agreements enabling coordinated investigation, evidence sharing, and perpetrator extradition.
	Victim Empowerment	Simplified Reporting Procedures	Single-point reporting system eliminating repetitive victim statements through cross-agency integration.
		Victim-Informed Technology Design	Including victim survivors in co-design of technologies and policies ensuring interventions address real needs.
		Trauma-Informed Support Services	Specialized counseling and medical services designed around victim psychology and recovery needs.
		Peer Support Networks	Structured connections between survivors enabling mutual support and information sharing.
		Victim Agency in Prosecution	Enhanced victim participation rights in investigation decisions and case outcomes.

Figure 6: Final version of qualitative codebook summarizing key themes and codes from the three workshop sessions. Codes are grouped under high-level categories: challenges and speculative design ideas. See Section 3.5 for more details.