
“I THOUGHT YOU WERE THE UNCENSORED PLACE”: NORMS, RULES, AND MODERATION IN AI-GENERATED SEXUAL CONTENT COMMUNITIES

Lucy Qin*
Georgetown University
lucy.qin@georgetown.edu

Jaron Mink*
Arizona State University
jaron.mink@asu.edu

Elissa M. Redmiles
Georgetown University
elissa.redmiles@georgetown.edu

ABSTRACT

As AI-generated sexual content (AIG-SC) is increasingly produced, online communities have emerged to support creators’ needs. To understand whether and how community governance attempts work to prevent abuse while supporting free expression, we interviewed 24 members and moderators of large AIG-SC online communities (10,000+ members) with stated rules against creating and sharing abusive content (e.g., AI-generated CSAM). Through in-depth interviews, we offer insight into: (1) how and why these communities form; (2) implicit community norms; (3) explicitly stated rules—and their operationalization via content moderation; and (4) tensions between community values and moderation that leave space for abusive behavior.

Our findings reveal a complex picture: while many creators and communities have personal boundaries against abuse, advice and resources for creating *any* form of AI-generated sexual content are accessible to users regardless of their intentions. Further complicating community moderation are norms that center anti-censorship and non-judgment, which leave moderators to justify their actions using the limits of the law and terms of service. We end by reflecting on the ways in which technical, community, and legal governance may most effectively mitigate the production of abusive content.

Content Warning: This paper discusses sexual violence, non-consensual intimate imagery, CSAM, and first-hand accounts of people who have perpetrated image-based sexual abuse with AI-generated imagery.
--

* **Both authors contributed equally** to this work, share first authorship, and reserve the right to adjust ordering on personal documents.

A version of this work appears in the proceedings of the 2026 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES).

1 Introduction

Like prior forms of technology-mediated sexual content [30], questions of whether AI-generated sexual content (AIG-SC) is produced or consumed ethically have become a focused societal concern. The public availability of retrainable open-weight models (e.g., SDXL, Flux), model repositories (e.g., CivitAI), and AI jailbreaks have enabled creators to generate virtually any form of sexual content [38, 52, 62, 92]. While some may legally create AIG-SC to explore sexual, creative, and technical interests [62], others use these same technologies to harm others in AI-generated non-consensual content (AIG-NCC¹) via non-consensual intimate imagery (AIG-NCII; Gibson et al. 32, Han et al. 35, Timmerman et al. 94, Umbach et al. 98) and child sexual abuse material (AIG-CSAM; Wei et al. 103).

In response, private and public institutions have taken measures to attempt to reduce AI-enabled sexual violence. Federal institutions and private institutions have appropriately taken down communities dedicated to enabling harm, such as the MrDeepFakes forum [35, 94] and Telegram channels [10, 95]. Similarly (and perhaps in light of irresponsible AI developer behavior; Conger et al. 17), AI developers have largely prohibited the generation of all sexual content, abusive or legal, within their platforms. While lauded by some, digital civil rights groups are increasingly concerned about whether lawful (sexual) expression is being censored and deplatformed [12, 64]. At the same time, others are weaponizing lawful expression to protect NCC-producing ‘nudeify’ apps [48].

In the wake of these rapid legal shifts and deplatforming of sexual content by mainstream AI-platforms, numerous peer-organized online AIG-SC communities have emerged. While many communities may clearly harbor deplatformed abusers [20], other large communities posture as refuges from censorship and for free sexual expression, yet are against AIG-NCC.

However, it is unknown whether, and if so how, these communities prevent and/or enable abuse. This is especially important given that there are no defenses currently available that guarantee AI-based image-generation technology cannot be used to generate AIG-NCII [14]. Prior work has found that community-supported norms and rules are effective deterrents against cyber-crimes, and are essential (alongside legal and technical interventions) within a comprehensive response to cyber-crimes [106]. While prior work has investigated what, how, and why *individual AIG-SC creators* make content [62], to our knowledge, none have analyzed *AIG-SC communities* and whether (and how) they support or discourage AIG-NCC creation.

To analyze behaviors supported within AIG-SC communities, we draw on analytical methods from organizational psychology and institutional economics. *Group formation* [37] finds that the process by which communities are created often determines group behaviors, composition, goals, and outcomes; however, it is unknown:

RQ1 How and why do AIG-SC communities form, and their members join?

Once groups are established, Ostrom’s institutional development and analysis framework [70] notes that *group (a) norms, (b) rules, and (c) governance* are essential in predicting and governing their behavior. However, within AIG-SC communities, it is unknown:

RQ2 What norms dictate interactions, and what beliefs are professed among members?

RQ3 What personal commitments/rules do (a) individuals and (b) communities use to govern their actions?

RQ4 What tensions, failures, or challenges exist within community-based AIG-SC governance?

To answer these questions, we conduct 24 in-depth interviews with members of three large (10k–50k+ members) AIG-SC Discord communities (two image-focused, one text-focused) that publicly prohibit AIG-NCC.

2 Related Work

AIG-SC. With the growing availability of AIG-SC generation services and resources on training customized AIG-SC models, it is now possible to create seemingly endless amounts of personalized sexual content via specialized AIG-SC websites [45], online forums [26], or even with assistance from other AIG-SC creators [62]. Through interviews with AIG-SC creators [62] and qualitative analysis of online AIG-SC applications [45, 46], multiple studies have examined how advancements in AI have enabled many to explore their sexual, creative, and technical interests, as well as enabled the production of AIG-NCII. Discussions and motivations for creating AIG-SC can be found intermingled with clear

¹We use the following acronyms to refer to different types of AI-generated content. AIG-NCII refers to AI-generated non-consensually created intimate imagery (sometimes referred to as “sexualized deepfakes” or “deepfake pornography”) that depicts a known individual’s likeness. AIG-CSAM refers to AI-generated child sexual abuse material. AIG-NCC refers to AI-generated non-consensual content, which we use as an umbrella term to include both AIG-NCII and AIG-CSAM.

intentions to make AIG-NCC of celebrities, work colleagues, friends, and ex-partners [26]. While some creators within these groups may openly condemn AIG-NCC, others rationalize such abuse [62].

Thus far, studies on the governance of AIG-SC spaces have been minimal; to our knowledge, only Lapointe et al. [46]’s analysis of AIG-SC terms of service, privacy policies, and related documents of AIG-SC (and AIG-NCC) generation and hosting sites has found that although policies against AIG-NCC exist, they largely serve as legal protection for site owners. While Lapointe et al. [46] also found that moderation appears to rely on user-reporting, no work has actually investigated how this moderation occurs in practice. Furthermore, no work has investigated whether AIG-SC creator communities establish normalized practices around their own content creation, the content creation of other members, or whether and how self-governance occurs.

AIG-NCC. Similar to AIG-SC, AIG-NCII and AIG-CSAM are becoming easier to create through nudification [32] and face-swap apps [23], model repositories [38, 54, 67] and AI companions [7]. AIG-NCC is even permeating mainstream platforms via listings on Fiverr [24] and AIG-NCC-capable chatbots on X [13, 107].

AIG-NCC creation has also been found to be supported through online communities. Prior to its shutdown, Han et al. [35] documented that the MrDeepFakes forum acted as a hub for information to create AIG-NCC, with skilled members teaching others. Although MrDeepFakes has since been shut down, AIG-NCC communities persist on platforms such as Reddit and 4chan [22, 60]. In response, researchers have investigated aspects pertaining to AIG-NCC, including: interventions to prevent AIG-NCC generation [3, 19], people’s preparation for and handling of AIG-NCC [73, 103], and offenders’ attitudes and behaviors [8, 29, 96, 97].

Moderation of Online Communities, Sexual Content, and AI Content. While several studies have investigated aspects relevant to AIG-SC communities, none explicitly address the unique legal, community-based, and safety concerns of this work. Several works have considered the difficulties volunteer moderators face within platforms generally, and have been centered around online forums [49], live streaming [109] and the emotional toll and burnout that often follows [83]. Other work has investigated how the moderation of content deemed as sexual disproportionately impacts vulnerable populations such as sex-workers [5, 93], queer communities [58, 89, 112], and vulnerable populations broadly [34]. Other work has more specifically focused on moderating *human-originating content from AI-generated content* [31, 51, 61], finding that such moderation is challenging and can lead to incorrect, biased decisions against real users [63]. However, no work investigates how moderators of AI-focused communities attempt to distinguish between AIG-NCC and AIG-SC, both of which are AI-generated and of a sexual nature.

3 Methodology

We interviewed 24 AIG-SC creators between April to August 2025 on (1) their background with sexual content and AI, (2) their creation process, (3) their participation within any communities centered on AIG-SC, (4) any norms/boundaries they had pertaining to content creation, and (5) their content or resource sharing practices. Procedures for ensuring safety for all researchers, participants, and potential victim-survivors are detailed in Section 8. Additional methodological details are documented in Appendix A.

Recruitment and Participants. We created a list of active Discord communities using relevant search terms that focus on generating AIG-SC and explicitly prohibit AIG-NCC.² We then reached out to the moderators to confirm their approval to interact with and recruit from their community. With approval, we distributed our recruitment information and responded to any requests for clarification from the community. Three Discord communities and their associated moderators participated. *Community A* (10k–20k members) and *Community B* (50k+ members) focused on visual content (e.g., images, videos); *Community C* (10k–20k members) supports users with text content (erotic fiction, role-playing). From each community, we interviewed between 2 to 14 members. Participants were offered \$40 USD.

Analysis. We conducted reflexive thematic analysis [6] and followed an inductive coding process. This began with each author reviewing a selection of interview transcriptions to explore themes within the data. They then met, discussed themes that emerged, merged their codebooks, and applied that codebook to three interview transcriptions that were triple-coded by all three authors. After discrepancies were resolved, this was then repeated with two more transcripts. The three authors then divided the remaining transcripts and independently coded them. We report on the thematic commonalities in what we observed.

Limitations. We do not claim generalizability of our participants or of all AIG-SC communities. We only selected communities with rules against AIG-NCC. Those who knowingly offend may be less likely to participate.

²Despite this, we prepared for, and encountered, AIG-NCC offenders (see Section 8).

Positionality. We conduct this work to enable prevention of sexual violence, while preserving individuals’ right to sexual expression. The team believes in a sex-positive approach to addressing and governing the production of sexual content. Our team contains members who hold nearly a decade of experience researching security, privacy, and safety concerns related to image-based sexual abuse, sexual content creators, and AIG-NCC. The team includes researchers who have experienced sexual abuse. None of the team actively participates in an AIG-SC community. We strongly believe in firm policies and protections against all forms of sexual violence, while also supporting individuals’ right to freedom of expression, including sexual autonomy.

4 How and Why Do AIG-SC Communities Form, and Their Members Join? (RQ1)

In asking AIG-SC moderators and members about how they became involved, we surface how/why their communities are created and why/how members join.

Discovering Communities. Search engines and general-purpose LLMs (e.g., ChatGPT) helped several participants discover AIG-SC online communities. For instance, C08 was curious about AIG-SC and naturally thought to *“Google like anything just to see like what is being done in the space.”* Additionally, participants such as A02 noted that posted tutorials and guides often linked to AIG-SC communities: *“I found a GitHub page that explained how to use [a video generation model] locally... I see [their community] was linked there in that repository”* For those focused on text-based roleplaying or erotic fiction generation, safety filters on general purpose LLMs (e.g., ChatGPT, Claude), led them to seek out AIG-SC communities. For example, C04 expressed that they were *“frustrated by forced restrictions that do not actually stop AI from being used in harmful ways.”*

Searching for and Exchanging Technical Knowledge. To support technical advice-sharing, two participants created AIG-SC communities on Reddit and Discord to centralize relevant information. For example, C01 had searched for reliable information in multiple sub-Reddits but found that they were mismanaged or contained spam. Afterwards, C01 *“decided to do it myself... and made sure to build [the community] in a way that everyone can have access to every tool they need to enjoy an uncensored experience with AI.”* Similarly, A01 shared the goal of cultivating spaces for open collaboration in order to improve tools and techniques for AIG-SC: *“my focus is on... getting people interested in the building, the research aspect of it, collaborating on best practices, exploring what the capabilities are”* (A01).

Members within these AIG-SC communities found that the quality and availability of technical advice on AIG content was higher in AIG-SC spaces than in other online communities that discuss generative AI, and even drew in those not initially interested in sexual content: *“I had joined many different Discord servers for people that are talented in this area. And it just so happened that the most talented people I’ve been able to find were in the Not Safe for Work space”* (B03). In order to find the best advice to support their AIG-SC experimentation, C11 also *“joined a lot of Discord groups.”* These AIG-SC communities were described as pushing technical boundaries. B07 notes that the community they participate in often tackles generation tasks not yet approached elsewhere, *“mixing concepts together, coming with more complex images, that is the thing that you see.”*

What started as technical help quickly turned into social bonding for several members: *“I would have them walk me through it... and then I’d forget everything and have them walk me through it again the next day... and it ended up being like over a month of consistently talking like every single night with these guys”* (B03). After receiving this assistance, participants reported feeling a responsibility to give back and assist others: *“I put effort into helping them with what I can do, because I like this kind of sharing the knowledge”* (B08). Some members, such as B07, described these AIG-SC spaces as *“a two-way helping hand,”* noting that from these experiences, their peers consisted of *“good people. [chuckles] They help everyone.”*

Beyond social connections, others felt that AIG-SC communities may also be useful in forging professional connections. For instance, C14 notes that some start-up ideas have emerged from Discord-based collaborations: *“The best thing that came out of it was the cool technical collaborations... maybe I’ll find something where I could get a job somehow through some of this”* (C14).

Rejection and Deplatforming. Participants created or joined AIG-SC communities on Discord because they were rejected elsewhere. They described varying tolerance towards AIG-SC in other peer-organized online communities. For example, B08 creates sexual content that depicts an anime character and *“the community around the character... they don’t like [AI]”*. When they previously shared their AI-generated outputs, *“people start to complain.”* Participants most often experienced rejection in communities with a tradition of artistic patronage: C07, an artist within the furry community³, mentioned hiding their use of AI as *“artists they tend to have a very negative viewpoint on AI, and if I*

³A fandom in which members depict alternative identities, often through commissioned art [2].

were to share that information that ‘hey, I enjoy creating AI images,’ a lot of them would be like, ‘okay, I don’t want to talk to you’...I can’t risk severing those connections.”

Conversely, more general online spaces (e.g., AI art communities) that participants frequented often prohibited sexual content. As such, AIG-SC communities served as a “*very niche special interest group*” (C14) for the convergence of these interests. Here, participants found consenting audiences with shared values: “*porn is usually something you keep to yourself but here you’re sharing it*” (B03). Even in offline environments, these communities served as one of the few, or only, places in which they felt comfortable discussing AIG-SC: “*I used to have a therapist... I was really open to her about everything else... but I never told her about [this]... I wasn’t sure if she would understand*” (C11).

However, even though these communities emphasize open-mindedness (see **Norm #3**) they still can be discriminatory: C14 mentioned that queer AIG-SC content was often marginalized, “*they’re into some extreme kink... but then they see two dicks, and they’re like, ‘Oh my God!’... it’s kind of funny how homophobia can present itself even in [the] extreme kink underground.*” For this reason, C14 and other queer members decided, “*let’s go to [a new AIG-SC community]... an all-M4M community.*”

Others found new AIG-SC communities when their prior ones were banned or moderated. “[*My prior community*] started an anti-not-safe-for-work crusade... Everybody scattered to other [AIG-SC] sites” (C04). Similarly, as the creator of Community A, A01 mentioned that they wrote an AIG-SC article and “*it got deleted from Reddit, which is what led to me creating the Discord.*” Others noted they “*didn’t like the way that [a prior community] was being moderated... All the old stuff that I posted got deleted... I was like, ‘Fuck that group!’*” (C14). Occasionally participants described deplatforming events triggered by policy changes due to public scrutiny:

“Nobody knows who, but someone went to TikTok, and showed... everything... people massive message[d] [the AI platform]... and [they] immediately [removed AIG-SC]... everybody was so pissed... a lot of them just went to [a different platform]” (C10).

5 Community Norms (RQ2)

Similar to existing works that identify social norms of online communities via qualitative interviews [27, 80, 85], we identify norms by analyzing behaviors participants report as being allowed, disallowed, and intentionally sanctioned by their communities, which indicate the existence of, and often define, upheld community norms [50, 68]. Through this, we find that AIG-SC communities often hold beliefs that (1) stand against perceived censorship, (2) support low-cost communal learning, and (3) reinforce judgment-free spaces, which are essential in advancing the expertise and infrastructure needed for AIG-SC creation.

Norm #1: “AI Needs to be Uncensored”. Participants wanted to create, and enable others to create, AIG-SC limited only by the bounds of their own imagination: “*the whole point of the Discord, of the community, is trying to make the generative AI models as diverse as possible, as customizable. Like, there’s no limitations*” (PB06). Corporate and regulatory governance over AI content generation were a large concern, which they viewed as censorship: “[*Companies*] censor. They have the power over what you can do, and what you cannot do” (B08). Such feelings were inflamed when credit card payment processors rescinded from CivitAI, a popular model-hosting platform, due to the prevalence of models specialized to create NCII [55]. Such action was viewed by several participants as AI censorship.

Those with text-based use cases often sought out AIG-SC communities after encountering safety filters on general LLMs (e.g., on ChatGPT or Claude): “*I got very annoyed at the sanitization of ChatGPT from filters, so I searched for NSFW AI communities*” (C06). Such participants were particularly sensitive toward censorship concerns. For instance, C04 expressed that they were “*frustrated by forced restrictions that do not actually stop AI from being used in harmful ways.*” Therefore, such participants intentionally sought out others who were interested in circumventing filters in order to enable sexual content creation: “*if you’re going to look for ways to get past [safety filters], you’re going to find out that other people have had these questions also and bit by bit, you end up in the sorts of places where people are wondering this in general*” (C03).

As a result of these censorship concerns, participants act to make AIG-SC communities havens for creative freedom with AI. People like PB06, who became interested in the AIG-SC communities after finding that it “*wasn’t some group that hell bent on making AI porn. It was just like, these guys are making [AI] uncensored*” believes that “*the whole purpose [of AI is that] you have the freedom to make what you want.*” Likewise, C04 describes that the AIG-SC community includes “*people who, like me, are frustrated by forced restrictions that do not actually stop AI from being used in harmful ways*” and their belief that “*if AI is really going to reach its full potential, where people who lack creativity but have great ideas are able to actually bring those ideas to reality... AI needs to be uncensored.*”

Limiting sexual content generation was perceived as the beginning of a slippery censorship slope: *“if you can block people from [making sexual content] then you can pretty consistently get [AI] to stop doing other stuff”* (C09). C09 draws parallels with prior forms of controversial art and creative expression that have been previously suppressed by institutions: *“Film history is full of like examples of people who made subversive things that were like horrible, or that were called various things but are still meaningful pieces of art.”* This discourse echoes academic analysis of AI-based censorship of artistic nudity [81], but also echoes arguments from AIG-NCC creators that the abusive content they generate is artistic [29, 35].

Norm #2: “Knowledge should ALWAYS be free”. Participants characterized AIG-SC community culture as driven by collaborative knowledge exchange of technical resources to advance the state-of-the-art. Although some members use AIG-SC communities to create products for monetization, free *information exchange* is a critical cultural value, as reflected by the amount of free, collectively-authored technical resources offered in AIG-SC communities and the availability of free peer assistance. Monetization was a divisive issue: as C07 described, *“Some people prefer it kind of not monetized, and others are more than happy to monetize the content and sell it for a profit. That’s often the divisiveness that I see...Most of the time, it’s over money.”* C01 noted that in another community *“people SOLD a jailbreak...that [behavior] goes against everything I believe and stand for...knowledge should ALWAYS be free.”*

Since participants valued free information exchange, concerns around monetization often stemmed from the belief that others were being scammed by private companies offering low-effort AIG-SC generators that relied upon the same free resources available to creators. This sentiment led C03 to provide the following characterization of those profiting from AIG-SC content generation products: *“They’re liars. They’re grifters trying to make money off of hooking up cheap models with simple instructions... And then getting people to pay... It is a terrible deal.”*

A shared cultural value around free information exchange also results in norms around free data collection and sharing. No participant mentioned paying for data (to e.g., fine-tune AI models) or obtaining consent to use it. As a whole, it is not clear where participants collected their data from, though in a few cases we observed that they collected content from social media and adult sites, such as *“Reddit and XHamster”* (A01). This is, unfortunately, in line with findings that social media and adult sites are being scraped nonconsensually to train AIG-SC models [16, 56]. Participants rarely engaged deeply with the ethics of nonconsensual data use. Rather, a few lamented that outsiders joined their community periodically to advocate against the theft of training data or business from artists. For example, B07 explains that sometimes *“there are some guys that just enter [the community] and say, ‘stop using this, this, this. Stealing artists’ works.”*

Norm #3: Don’t Judge, “Don’t Ask, Don’t Tell”. Given the stigma around sexual content and the ostracization participants may have received (see Section 4), they expressed a strong desire for nonjudgmental spaces that allow them to discuss and share the content they produced freely: *“You’re doing way more harm to this world by stifling sexual expression... sexual stimulation is a form of self-care”* (C12). As noted by several moderators, their communities were built upon norms of non-judgment that were often essential for encouraging the community participation required to develop AIG-SC creation expertise: *“we try to set up a place for people to learn how to use the tech so they can explore their niches and for people to share their art without fear of judgment”* (B01). This nonjudgmental spirit enables members to find interested audiences and explore sexual niches (e.g., content relating to a specific kink). For example, PB06 noted that, *“there’s all these people contributing to the knowledge of how to make photorealistic cat girls.”*⁴ While members may not all share the same sexual interests, *“there’s different people into different stuff. But it is generally pretty supportive”* (C14). As B03 described, *“it’s people genuinely sharing what they’re into and that might be shocking to some”* but *“people are really open-minded. And that’s actually really nice to just have people you can share whatever with, especially stuff that’s so intimate like this.”* Accordingly, participants tended to deliberately ignore others’ implied actions, as long as the resulting content was not shared: C09 notes that *“when it comes to like stuff that we ourselves generate it’s usually kind of said ‘don’t ask, don’t tell’”;* however, if content is shared that is *“very obviously undefendable, we will just complain about it and then [a moderator] bans them immediately.”* When discussing content with darker, more violent themes, B04 commented that, *“a lot of people make some make some pretty gruesome stuff. [laughs] And they don’t share. I’m pretty sure there’s been a lot of skeletons in other people’s closets.”*

Several participants, such as C05, believed that as long as the content is private, it does not matter what is generated: *“The community in general thinks that whatever you want the AI to write is okay, just keep it to yourself.”* Deliberate ignorance and belief that private generation does not harm others contribute to rationalizations of harm as exemplified by C04 who believed that because content can be generated privately, it is *“not harmful to anyone else.”* This was a recurring sentiment expressed by others, *“AI in of itself is not harmful or hurtful. It’s the people that use it”* (C07).

⁴Cat girls are a common archetype in anime, typically depicting a human woman with cat-like features (e.g., cat ears).

In combination with other norms that center creative freedom against censorship and free information exchange, the “*don’t ask, don’t tell*” mentality can lead to notions of futility around combating harm. B08 vocalized that their country has had “*a lot of incidents lately of students making naked images of their other female classmates... Or people making underage content... I don’t believe we have actual control over that. It’s like the side effect of creating this technology, and we will need to learn how to live with that.*”

6 Commitments and Rules (RQ3)

The availability of community-based upskilling, combined with open models and accessible fine-tuning techniques, enables highly customized content to be created with few safeguards, “*You cannot control what people create. There are free models, and people can create whatever they want with that. There’s no way to stop that*” (A02). To guide what content they create and share, creators negotiate the community norm of uncensored content generation against their own personal commitments [84], and explicit community rules [70].

6.1 Personal Commitments

While some participants describe their personal commitments as “*way less restrictive*” (B07) than what is allowed in their communities, others personally opted not to share “*anything that I feel is morally wrong... none of it’s illegal [but] if there’s anything I feel like, ‘Okay maybe it’s a step too far in the content here,’ I don’t share it.*” (B04).

When asked about their boundaries, most participants explicitly brought up a personal commitment not to generate underage content (either text or images). Some, but fewer, mentioned commitments against creating NCII. To avoid creating NCII, B07 focused on creating anime AIG-SC models without using images depicting humans, stating that, “*I will not train on someone without them knowing that I’m doing this.... I don’t want to get sued.*” However, a few individuals excluded content depicting celebrities from NCII, viewing these individuals as more implicitly consenting.

In a few cases participants who take commissions received requests for NCII and expressed a clear boundary: “*I’ve gotten requests for commissions for deepfake content, which I’ve absolutely refused... because of my own ethical and moral compass. That’s really where I draw the line*” (B05). Avoiding NCII creation in practice means potentially incorporating more vetting. As further described by B05, they have also been approached by people who want to generate content depicting themselves: “*[An OnlyFans Creator] wanted me to create, AI content of [herself]... she sent me a bunch of reference images*”. This places responsibility on creators to verify that the images actually depict the individual making the request. Further research is needed on if and how they perform such verification.

While most participants shared a personal commitment not to generate underage content, AIG-SC communities, in general, were described as containing individuals with varying personal commitments. Personal commitments do not necessarily translate into governance of others’ behavior. As B08 explains: “*I don’t do what I don’t like... [such as] underage characters. I myself believe it’s disgusting, to be honest. [But] for me, it’s not like if some people I know like it, I will confront them telling it’s bad... that’s their problem, not mine.*” Explicit rules, along with moderation, are necessary to establish collective baselines: “*communities often have some kind of ethical guidelines they establish or topics that they request you stay on when posting. Super common*” (B02). In comparison to spaces that conventionally lack moderation, such as 4chan or Telegram, “*Discord is much kinder as a rule; self expression will be more regulated to some degree. Either by the self because it’s not anonymous or by the community because they don’t tolerate certain types of behavior*” (B02). However, the level of depth in community rules and stringency of enforcement vary.

6.2 Community Rules

Within all three communities, there are explicit, public community rules (e.g., posted in a designated channel) that are summarized in Table 4 in Appendix B.2. We provide this analysis to give a snapshot of the types of rules that exist within AIG-SC communities and to supplement participant narratives. This does not serve as an evaluation of these communities, or AIG-SC communities in general, especially since rules are often in flux and adapt over time: “*I just put up new rules just earlier today... we’re going to see the server kind of evolve*” (A01).

AI-generated Output Sharing. The vast majority of rules across communities focused on AI-generated output sharing. Text content is less commonly shared for a variety of reasons (see Appendix B.1). Therefore, rules often pertained to visual content. Notably, rules are lacking for asset creation and sharing (see discussion in Section 6.2).

Age: Without stringent moderation, a few participants noted that underage content is, in general, “*the most popular theme to develop*” (B08) across different online spaces (including those not on Discord). Therefore, some participants emphasized the importance of explicit rules and active moderation of underage content:

“[AI] opens the possibility for anyone to create basically anything. [People have] weird fantasies and stuff...that’s [mostly] fine... But I have seen a lot of...the child fantasies or pedophilia... it was really, really concerning and really weird seeing images, like, of pictures of girls, small girls... It’s not AI’s fault... if [people] have a tool to do something, they are going to do it. So, yeah, I think that it’s better to have a little bit of control on what, at least publicly, [is] available for people to see.” (A02)

In line with participants’ personal commitments, all three communities had public rules against sharing content depicting minors. Participants noted that the rule against underage content superseded **Norms 1 and 3**. For example, B03 explains, *“I might not want to see these fat furies with big dicks, or something like that... But [the community is] going to be like, ‘Okay, that’s your own right to generate that.’ [But the] child stuff... people can agree on, ‘Okay, we’re not going to generate this.’”* Similarly, C03 states that *“obviously, anything sexually charged can be a bit sensitive”* and they try to be accepting and *“to understand that every person is different...”,* but *“there are some things that definitely do cross lines and we’ll tell people [in my community], ‘No, this is not happening, this is not the place for you.’”* However, given the inherent ambiguity of AI-generated content, there can be challenges in distinguishing what “crosses a line” as we discuss further in Section 7.2.

NCII: Similar to participants’ personal commitments, NCII was the next most common content rule mentioned: *“just don’t make anything featuring real human beings or children and I think you’re fine legally”* (B02). To actualize this, both *Community A* and *Community B* had rules against sharing content depicting real people.

Assets. In contrast to the development of stringent rules on sharing AI-generated outputs, there is a noticeable gap in rules around asset sharing.

Tools Enabling CSAM: Generally, there were no stated rules about sharing tools and assets that enable the creation of underage content. In combination with **Norm #2**’s emphasis on free information exchange (see Section 5), the absence of such rules may inadvertently enable the creation of AIG-CSAM and AIG-NCII. *Community C* was the only one that had a rule against sharing assets that could enable underage content. Since they focused on text-based use cases, their rules included sharing jailbreaks or prompts that could be used to generate underage content. Community members enforce this by refusing assistance with related prompts and by directly confronting those who violate the rules, *“There are a whole bunch of people now and then who occasionally enter... you will question those people and figure out that, ‘Huh, these things you’re trying to have it write, these people are wildly underage. We do not approve of this, get out of here’”* (C03). However, this mostly applies to overt requests since members cannot dictate how others use jailbreaks once they are shared. To avoid jailbreaks that are overly permissive as to allow underage content, *Community C* created their own jailbreaks. Unlike sites like 4chan that C09 characterized as *“completely unfiltered... and there’s nothing that you can do about that”*, this community bans those who violate their rules *“as fast as possible”*.

Data Sharing: *Community A* was the only community that had a stated rule against directly sharing visual data that could be used for training (e.g., posting data in the Discord server). However, as the server also allows links to datasets and even self-hosts *“huge files, tons of videos, tons of images, multi-modal sort of stuff... there’s options where nobody’s going to tell us to take it down”* (A01) in order to evade legal and platform regulation, it casts doubt on whether such rules are genuinely enforced.

7 Enforcement and Governance Gaps (RQ4)

Three participants (A01, B01, C01) moderated communities and provided significant insight into the mechanisms of moderation within their communities, supplementing community members’ narratives about moderator actions. While our discussion focuses on the three communities we recruited from, participants were often active on other Discord servers, where they observed varying levels of moderation. For example, B01 observed that on *“other servers and other parts of the internet [moderation is] done very poorly or not at all”*, which they describe as an issue not unique to AIG-SC and, *“just the reality of our internet.”*

7.1 Enforcement

Community Owners and Moderators. Given the size of the communities and how quickly they may grow, the amount of content shared scales quickly. Moderation can be taxing *“when you have that much activity with that many people, it’s very hard to keep track of everybody. So, that was the constant struggle”* (B01). They further added that because their community has a large number of members: *“you just need people [moderating] 24/7, like you can’t go ten minutes, I would say, without viewing the server, without something being posted that needs to be checked.”*

Some chose to only allow output sharing in designated channels, which may help limit the number of channels that need to be more stringently monitored. *Community C* leveraged Discord's spoiler tag feature, which blurs content by default, to reduce the visibility of shared content. Users must opt into viewing the content: *"We do not shove our kinks at people... it's always behind spoilers... [members] have to click on it to see it. It's their choice."* (C04)

Community Members. *Community B* and *Community C* explicitly ask members to report others who are breaking the rules, particularly given the borderless nature of online communities. As B01 describes, *"There's like millions of celebrities... maybe someone will recognize someone from another country."* Proactive reports also help moderators stay informed as communities expand:

"We did have a few instances where someone will...go around asking [for NCII]... luckily our most active members... they'll be like, 'Hey, we're just letting you know that so-and-so sent me a DM saying they'll pay me for basically deepfaking someone.' And so we'll just ban the person once [the reporter] post[s] proof" (B01).

Community members often have vested interests in sustaining online spaces and safeguarding them (e.g., from deplatforming). *Community A* does not ask members to make reports, and perhaps relatedly, is the only community to note that users retain full liability for their actions. However, moderator A01 observed that in this community, non-moderator members enforce rules and norms *"more seriously than I even have... People see this as a valuable place, and they want it to be protected."* Therefore, community members may proactively confront one another about behavior that violates stated rules. As described in Section 6.2, within C03's text-based community, members will *"gang up and go 'get out of here'"* if they see others asking for advice on creating content involving underage individuals.

Self-Policing. As discussed in Section 6.1, individual members may vary in their personal commitments regarding what content they will create, share, or view. If members create content that does not violate their own personal commitments but are outside the bounds of community rules, they may refrain from sharing it to preserve the community: *"If I can't post it, I'll just not post it... it's less about keeping the community as clean as they want. It's more about not getting banned on Discord"* (PB06). In a few cases, participants were warned about their content and complied. For example, B08 refrained from further posting content because, *"they told me that the character looks... underage. And I was like, 'I'm making my best not to look, because the character is actually not underage'... [with] more cartoonish things, it's very difficult to tell the age... So, I stopped [posting]."*

7.2 Governance Gaps

Ambiguity in Enforcement. AI-generated outputs introduce inherent ambiguity to moderation decision-making because, *"when you generate something, it's not like you know their age because they don't exist. You just have to guess by looks and that can always be very, very subjective depending on where you're from in this world"* (B01). Ambiguity around age, particularly in animated content is also a challenge that platforms are contending with [75].

Moderators often relied upon or partnered with community members to adjudicate such cases. As C01 explains, *"it's a bit more complex to judge [images], so it's case by case... I also trust my community to judge if they think a piece of art represent a underage character."* B01 described how they not only moderated the appearance of the individual in an image, but its context and implications, *"like school settings or school uniforms, anything that might suggest a younger age... we always are on the side of caution because we have no idea [what] the person [at the hosting platform] reviewing the server [will think]."* In response to the same rule, C14 felt "school settings" was an imperfect barometer: *"'Oh, your image had a chalkboard in the background?' No!... that just seems like kind of, overkill to me."*

Content that depicts fictional anime or cartoon characters is further complicated by different cultural norms. For example, B08 explains that while they personally find underage anime characters to be *"disgusting"*, there are cultural and legal nuances to what content is permissible in different countries: *"In United States, there are some pages you can upload a full naked character and in Japan, you need to censor some parts of it, because of legislation regulations...There are other things that are clearly illegal in America that are not illegal in Japan⁵... it's very delicate."*

NCII also presents challenges for moderation since, *"you don't know if this is purely AI-generated, or if it's a person. If it is a person, is it just a face swap, or is this their body, is this an original video, is this an AI video. Do they even have consent? Yeah, we don't know"* (A01). This challenge is further exacerbated by the rapid advancement of visual content generation techniques that make it difficult to distinguish AI-generated content from human-produced content.

Legitimizing Enforcement: Laws and Terms of Service. The AIG-SC communities our participants are involved in are typically created by a single individual or a small group. In their early stages, communities tend to have fewer

⁵A 1999 Japanese law on child pornography which does not apply to AIG-CSAM [88].

formalized rules and might be *“fairly lenient because you’re not going to assume the worst of someone just because they’re acting up a bit”* (C03). Within smaller communities, it may also be easier to establish *implicit norms* that do not need to be codified. As membership grows, it becomes more difficult to monitor the volume of content. Growth can bring greater visibility to the community’s activities as well, increasing potential legal risk for community owners. In B01’s experience, *“We started off just a small Discord... but it kept growing... It was very lax in the beginning”*. As they grew, they developed *“strong rules”* and *“tend to be a little bit more strict... that’s just out of safety and precaution for us.”*

When creating and enforcing rules, moderators are motivated by their own commitments, their members, and by concerns around deplatforming—*“it’s more about not getting banned on Discord”* (PB06)—as well as personal liability: *“if something happens... [Discord is not] going to look at the users, but the moderation team, and the person running the [community]”* (B01).

In an environment that prioritizes free expression, nonjudgment, and reducing stigma, several community members mentioned that it can be difficult to speak up about content that makes them uncomfortable, until it undeniably crosses a line (e.g., CSAM); even then, they may simply avoid it: *“I understand [it’s others’] kink or whatever, but it’s certainly not something that I’m into at all... that’s something that over and over again I’m just X-ing out of, or hiding from my page”* (C11). Moderator A01 explains that if they overmoderate, their members will push back, saying *“I thought you were the uncensored place’... so I want to take the stance of legal[ality] is where I draw the line”*.

Laws regarding CSAM offer a “clear dividing line:” *“you can absolutely go to jail if you are spreading that content in a manner to encourage others to take part in, say, pedophilia. It’s very, very, very clear dividing line... it’s very cut and dry”* (C04). *Community B* included warnings about complying with subpoenas and reporting to the U.S. National Center for Missing and Exploited Children in the United States, which operates a hotline for reporting child sexual abuse material [66]. Conversely, there are fewer established laws against creating NCII. Therefore, A01 found it harder to explain enforcement decisions demonstrating the importance of legal precedent in strengthening moderation decisions. For example, they have encountered cases where members *“shared a video and they mentioned it’s of their ex-girlfriend.”* Although they had a discussion with the member, they *“don’t know that there’s a hard legal line on that, so I’m not even sure where to go.”*

While laws may differ depending on the country, B01 feels that *“mostly everyone who runs [the Discord server] is located either in the US or other Western countries and so we have to play by those rules at the end of the day.”* Community members reside in any part of the world and may not agree since the *“law on underage content can vary and can be legal as long as it’s fictional in some country”* (C01). Therefore, *“sometimes they’ll open a ticket or they’ll complain”* (B01) but moderators can defend their decisions by instead pointing toward Discord’s terms of service, which: *“say no underage sexual depiction so I go with that, don’t want my server to get nuked”* (C01). In a different server, C03 also noted that *“the Discord terms of service would also tell us to kick people into underage content out... which is, again, fine in my view... go find some other people to bother.”* The moderator of *Community A* instead shared that beyond the bounds of the law in their jurisdiction, they, *“don’t even really care about Discord’s rules... I would rather make the community happy.”* In case their community is banned, *Community A* has already created a backup Discord server and explains *“if we get banned, we have to go find somewhere else”* (A01).

8 Discussion

Across our interviews, we observed a cultural ethos toward free knowledge exchange and uncensored AI generation. In part due to perceived rejection from other communities, participants desired nonjudgmental spaces to discuss AIG-SC creation. While these cultural values enable creative and sexual exploration, they also create challenging tensions for moderation. AIG-SC communities attract a range of individuals, including those who strongly oppose abusive content alongside some who use community resources to produce it. Since AIG-SC communities can be surfaced through simple web searches and are at lower risk of being deplatformed, they are highly visible and accessible. Their governance, which attempts to prohibit abusive behavior while upholding their values, may be particularly important towards discouraging AIG-NCC offenses. As such, we believe AIG-SC communities are important sites of intervention; we now provide several recommendations for doing so.

Moderation Needs Assistance. Supporting sexual expression and condemning sexual violence are not contradictory actions; sex-positive education and perspectives can serve to mitigate actions and beliefs that perpetrate sexual violence [11, 74, 87, 104, 105], and help victim-survivors process trauma [4, 71]. However, clearly distinguishing what content is or is not abusive (particularly with little context) can be challenging and is an issue moderators face (Section 7.2). These moderation challenges mirror those encountered by platforms when attempting to prevent AIG-NCC generation at the model level, or the sharing of it on social platforms. Automated detectors of children in images have known inaccuracies [19] and distinguishing AIG-NCC from AIG-SC requires not nudity detection

but rather identifying the likenesses of real people in generated content, which remains an unsolved problem with privacy implications. Beyond challenges in accurate content classification, community moderators report having to review an overwhelming amount of content, and handle the emotional toll caused by repeatedly viewing content flagged as AIG-NCC (Section 7). Added to these domain-specific challenges are the general challenges faced by Discord moderators: that moderation relies on volunteer labor [57] and encompasses text, voice, and video; the latter are more ephemeral and fundamentally more difficult to capture evidence for enforcement from [43].

In light of these challenges, even members who hold strong personal commitments against AIG-NCC can begin to view this content as an inevitability (Section 5, *Norm #3*). This minimizes harm as a byproduct of AI development and positions the technologies as neutral. This same type of justification is used by communities that condone AIG-NCC, such as commenters in face-swapping repositories: “*like any technology, it can be used for good, or it can be abused*” [67]. While non-abusive use cases for sexual and creative exploration exist [62, 81, 90], complicating the narrative is the fact that those who create AIG-NCC have been known to co-opt these rationales [29, 35], which can make it difficult to discern members’ true motivations.

Given both the importance and difficulty of providing clear, justified, and correct moderation decisions in these communities in an AI industry that seeks to enable generation of “NSFW stuff for your personal use... but not do stuff like make deepfakes” [1], we recommend that future work investigate how to build sociotechnical processes that work for moderators who lack corporate moderation resources.

Prevention Focuses on Outputs Not Infrastructure. In the AIG-SC communities we study, governance against AIG-NCC focuses nearly exclusively on outputs, not the assets and behaviors that enable those outputs (text-focused *Community C* was the exception; Section 6.2). This is a significant governance gap, given that these communities serve as key hubs, amassing resources and exchanging knowledge to enable *any* type of sexual content creation (Section 5).

Data is collected and exchanged in accordance with the communities’ norms of free knowledge (*Norm #2*); *Community A* even goes so far to adopt practices to minimize possible legal repercussions. Several participants openly discussed non-consensually scraping and using data from adult websites and forums, and no participants discussed proactive measures to verify and remove content that could be illegal or violating (e.g., CSAM). These are reminiscent of general ML practices, with CSAM discovered in LAION [92] and NudeNet [53] datasets, and NCII in nude-image datasets [16].

Models are shared, bought, and traded without a clear understanding of what the model was trained to generate, or what it was trained on. In peer interactions, none of the participants mentioned vetting others in the community to ensure that they were not engaging in AIG-NCC prior to sharing advice; nor did those taking commissions discuss verifying authenticity from those claiming they were commissioning images depicting themselves. Rather, the community “*don’t ask, don’t tell*” (*Norm #3*) meant giving other members the benefit of the doubt and avoiding responsibility. In combination with *Norm #2* of free information sharing, existing governance in AIG-SC communities can inadvertently enable AIG-NCC through shared knowledge, peer assistance, and shared technical assets (e.g., datasets, custom-created tools). Indeed, we find significant resemblance between the resource infrastructure supporting AIG-SC communities and those supporting AIG-NCC [35, 60].

Regulate AIG-NCC Creation, Commission, and Tech. Although AIG-SC communities are frequented by individuals across the globe, regulation still plays a critical role in establishing a baseline for moderation and supporting enforcement decisions. While the impacts of regulation appear not to be uniform across AIG-SC spaces, particularly within AIG-NCII-supporting spaces such as 4chan [21], our findings suggest that among those who attempt to create AIG-SC in accordance with their own commitments and community norms, regulations provide impetus and justification for moderators to establish rules that echo regulation, justify the removal of problematic content, and reinforce norms (Section 6). Legally influenced rules and their enforcement may be particularly important in communities. First, it may be easier for impressionable newcomers to the space to justify joining such communities rather than more abuse-centric ones. Second, moderators must be responsive to their community; if decisions are not appropriately communicated or justified, members may view these actions as at odds with held norms of uncensored creative freedom and open-mindedness (Section 5, *Norm 1–3*). As members are not bound to any community, moderation perceived as improper may lead to migration towards less-regulated spaces that encourage abusive behavior.

Unfortunately, AIG-NCC regulation remains nascent and inconsistent across jurisdictions. In the U.S., CSAM laws [28] criminalize both the production and distribution of AIG-CSAM, while the TAKE IT DOWN Act [91] only criminalizes the nonconsensual distribution of AIG-NCII. Similarly, we find that posted community rules primarily focus on preventing CSAM outputs or, in one case, tools that enable CSAM generation (Section 6.2). Other regions have taken different approaches, including calls to ban the marketing of technology that enables AIG-NCC in the European Union [77], criminalizing the viewing and possession of AIG-NCC in South Korea [33, 110], and using copyright law to ban any unauthorized use of an individual’s likeness in Denmark [9]. However, these laws largely focus on the generated outputs themselves, without consideration for the broader ecosystem of assets that enable AIG-NCC creation.

As one example, the UK recently amended an existing law to criminalize requests to commission AIG-NCII, regardless of whether creation takes place in the country. This accounts for offenders who may purposefully outsource creation to countries with little (or no) regulation [41, 72]. Without stringent regulation, offenders may believe there are no consequences to creating AIG-NCII, which helps rationalize their actions [29]. Given our findings that regulation has a substantial downstream impact on platform policies, community governance, and the beliefs and actions of individual creators, we echo calls for regulation against the broader AIG-NCC enabling ecosystem.

A Public Health Approach to AIG-NCC. As scholars have previously identified, addressing sexual violence and child abuse requires a public health approach [47, 59] that invests in the development of interventions before harm occurs (*primary prevention*) and immediately after harm has occurred (*secondary prevention*). Despite their effectiveness at reducing overall offenses, such interventions are still limited for AIG-NCC.

Primary preventions for AIG-NCC should consider the general populations that interact with these AI systems. These could include developing and embedding computer science curricula that incorporate sex-positive education on consent proven to reduce offenses and offense-enabling beliefs [74]. At a minimum, discussions on the ethics of non-consensual data collection and use, the dual use of generative AI, and the harms of AIG-NCC should be embedded. This may help create and reinforce norms that reduce cyber-deviant behavior and crime [106].

Primary and secondary interventions should more intentionally consider how to intervene with potential offenders (e.g., through visible behaviors). While there is emerging work on deterrence messaging in the context of NCII [102], deterrence messages for those seeking CSAM have already been integrated into online platforms (e.g., Pornhub, Google search), which early evidence suggests is effective at reducing abuse-seeking searches [78, 82]. When messages are further combined with links to anonymous services, evidence suggests that a non-trivial number of (potential) offenders may stop and, in some cases, seek therapeutic help [82, 99].

Efforts are also needed to add friction to AIG-NCC creation. Experts should work to develop and evaluate effective peer-based deterrence messages to support moderation within AIG-SC communities. AI developers should consider implementing deterrence messages for prompts that appear to request, or seek advice on creating, AIG-NCC. We urge that messages be developed in partnership with experts, as subtle variations can impact efficacy [15].

Beyond single-turn deterrence messaging, comprehensive programs are needed to address AIG-NCC. Educational programs shaping sexual violence-related attitudes and norms are recommended by the UN [100] and services such as peer support groups, self-help resources, anonymous helplines, and clinician-developed programs have been developed for CSAM-specific offenders and people at risk of offending [40, 76]. It is worth considering how similar services could be offered to reduce AIG-NCC. It may also be useful to consider these services as alternatives to company-imposed punitive threats or deplatforming, which, may instead *increase offenses* by moving those at-risk of offending to more extreme sects of the internet [20].

Ethical Considerations

Throughout this research, the research team engaged in ongoing discussions on how to ensure this work could be done ethically and responsibly. This was considered both through our research methods with participating stakeholders (i.e., the study participants and interviewing researchers), and the expected impacts this work may have, including how it may impact non-abusive sexual content creators, those impacted by labor displacement and intellectual or labor theft (e.g., from sex workers and content creators), and victims-survivors of AIG-NCC. We now discuss the procedures we followed to minimize risks. All procedures were approved by the IRB of each author's institution.

Research Procedures

While the primary focus of our research concerned how members and moderators interacted and governed themselves in AIG-SC communities with explicit rules against the sharing of AIG-NCC (i.e., AIG-NCII and CSAM), we prepared for the possibility that members we interviewed may be offenders of sexual violence. To account for this, we created a set of procedures for ensuring the entitled (but limited) protection of all research subjects, as well as procedures we would follow if a participant revealed that they had perpetrated, or intended to perpetrate, sexual violence.

General Procedures For Participant Protection. For the general participant procedure, we follow best practices to ensure participant protection. All participants were recruited from large communities of over 10,000 members. Our research team received permission from moderators explicitly before interacting with any members or posting content within their community. Every participant was provided instructions for how to email and conduct the interview anonymously. Participants could choose to be interviewed over video, audio, or text. Prior to recording, each participant

was reminded of the study's purpose and asked to reaffirm consent before being recorded. Participants could end the interview at any point, skip any question, or retract their data after the interview, though none did.

Procedures for Disclosures of Sexual Violence. While we intentionally limited our recruitment to members of communities that explicitly prohibit the sharing of AIG-NCC, we prepared procedures in case a participant disclosed AIG-NCC creation during an interview. To build our procedures, we referenced two key sources. *First*, established PIs and researchers who have experience either working in a sub-domain in which participant retaliation was a consideration, or working in a field such as psychology and/or criminology in which established norms for conducting qualitative research with offenders of sexual violence existed. *Second*, we extensively utilized the recommendations from established guidelines for interviewing offenders of sexual violence. Specifically, we considered how best to protect researchers from harm during these interactions and minimize harm to current or future victim-survivors.

Despite our recruitment methodology, three participants disclosed that they had created AIG-NCII during the course of their interview. We now discuss the procedures we have prepared for handling such disclosures.

Limited Confidentiality: To prioritize the current and future safety of victim-survivors, we follow established guidelines [18, 39, 42, 86] and implement limited confidentiality within our study procedures and consent process. If a participant discloses a clearly identifiable criminal offense in which they (1) create CSAM, or (2) publicly upload AIG-NCC of an identifiable victim, or state an intention of imminently doing so, we would end the interview (if active), report the activity to NCMEC cybertip line [65] (for CSAM) and a platform authority (if publicly uploaded). This limited confidentiality was established within our approved IRB procedures and similarly disclosed when participants' consent was obtained. Following best practices as researchers [18], we do not act as law enforcement and do not attempt to elicit more details than necessary to complete our research, and we do not attempt to elicit information that would require breaching confidentiality. Ultimately, we were required to breach confidentiality for one participant who sent the interviewer their public account over chat during the interview. After the interview concluded, publicly posted NCII of an identifiable adult was discovered, and that content was reported to the platform and removed. After our study concluded, moderators from the participating communities were informed of these results, including the existence of offenders of sexual violence who may have been recruited from their community.

Preventing Condonement of Sexual Violence: To prevent normalization of AIG-NCC behaviors, we follow established guidelines [39, 42], and ensure that interviewers do not condone acts of sexual violence either through verbal or non-verbal cues to any participant. To protect researchers and prevent confrontation, interviewers did not provide any opinions on offending behavior, but did question the appropriateness of such behavior if minimized by the participant [42]. As a result of our interview, participants appeared to be aware that the creation and distribution of the AIG-NCC is unacceptable. PC13, who created AIG-NCII of a public figure, later denied ever using AI to make sexual content; another moderator who did not engage in AIG-NCC added significantly more restrictions to their communities' rules, including those relating to AIG-NCC.

Protection of Researchers: All researchers on this study were aware that the work carries a non-negligible risk and openly discussed this; in addition to being a potential target of AIG-NCC, potential community backlash is unfortunately a real possibility [25]. All researchers were provided access to a mental health consultation throughout the project. As this was the team's first time interacting with these communities, and out of an abundance of caution, we attempted to minimize power dynamics by not including any persons in degree-granting programs (e.g., PhD, Masters, Undergraduate). Our team fully believes that including students is necessary for academic mentorship and establishment of norms but made this exclusion out of extreme caution in interacting with an unknown community. We have since established trust with moderators in AIG-SC online communities and are conducting follow-up work that includes students.

Acknowledgments

This work was supported in part by NSF Award 2513313. We thank Grace Brigham, Bernardo B. P. Medeiros, and Galen Weld for their detailed feedback on prior versions of this work. We thank Filippo Sharevski, Gianluca Stringhini, and Sharon Wang for providing advice on researcher and participant protection.

References

- [1] Sam (as u/samaltman) Altman and OpenAI team. 2024. r/ChatGPT is hosting a Q&A with OpenAI's CEO Sam Altman. https://old.reddit.com/r/ChatGPT/comments/1coumbd/rchatgpt_is_hosting_a_qa_with_openais_ceo_sam/
- [2] Jessica R Austin. 2023. *Identity construction in the furry fandom*. Ph. D. Dissertation. Anglia Ruskin Research Online (ARRO).

- [3] Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. 2024. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution. In *Proc. of CCS*.
- [4] Linda R Baggett, Ethan Eisen, Sara Gonzalez-Rivas, Lacy A Olson, Rebecca P Cameron, and Linda R Mona. 2017. Sex-positive assessment and treatment among female trauma survivors. *Journal of clinical psychology* 73, 8 (2017), 965–974.
- [5] Danielle Blunt and Zahra Stardust. 2021. Automating whorephobia: sex, technology and the violence of deplatforming: an interview with hacking/hustling. *Porn Studies* 8, 4 (2021), 350–366.
- [6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* (2006).
- [7] Natalie Grace Brigham, Lucy Qin, and Tadayoshi Kohno. 2026. Examining Risks in the AI Companion Application Ecosystem. doi:10.48550/arXiv.2603.13620 arXiv:2603.13620 [cs].
- [8] 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 *Proc. of SOUPS*.
- [9] Miranda Bryant. 2025. Denmark to tackle deepfakes by giving people copyright to their own features. Published: The Guardian.
- [10] Matt Burgess. 2024. Millions of People Are Using Abusive AI ‘Nudify’ Bots on Telegram. <https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/>
- [11] Moira Carmody. 2005. Ethical erotics: Reconceptualizing anti-rape education. *Sexualities* (2005).
- [12] Center for Democracy & Technology. 2025. TAKE IT DOWN Act Sign-On Letter: Concerns Regarding S. 146. https://cdt.org/wp-content/uploads/2025/02/TAKE-IT-DOWN-Sign-On-Letter_21225.pdf
- [13] Kelvin Chan. 2026. Musk’s Grok chatbot faces EU privacy investigation over sexualized deepfake images. *PBS* (Feb. 2026).
- [14] Bilva Chandra, Jesse Dunietz, Kathleen Roberts, Yooyoung Lee, Peter Fontana, and George Awad. 2024. *Reducing Risks Posed by Synthetic Content An Overview of Technical Approaches to Digital Content Transparency*. Technical Report. NIST Trustworthy and Responsible AI.
- [15] Caoilte Ó Ciardha, Joel Scanlan, Tegan Insoll, Juha Nurmi, and Nina Vaaranen-Valkonen. 2026. Warning Message Content Increases Help Seeking in a Large-Scale Dark Web CSAM Intervention. doi:10.48550/ARXIV.2606.06417 Version Number: 2.
- [16] Princessa Cintaqia, Arshia Arya, Elissa Redmiles, Deepak Kumar, Allison McDonald, and Lucy Qin. 2026. Stop the Nonconsensual Use of Nude Images in Research. In *Proc. of NeurIPS*.
- [17] Kate Conger, Dylan Freedman, and Stuart A. Thompson. 2026. Elon Musk’s Grok A.I. Chatbot Made Millions of Sexualized Images, New Estimates Show. <https://www.nytimes.com/2026/01/22/technology/grok-x-ai-elon-musk-deepfakes.html>
- [18] Malcolm Cowburn. 2005. Confidentiality and public protection: Ethical dilemmas in qualitative research with adult male sex offenders. *Journal of sexual aggression* 11, 1 (2005), 49–63.
- [19] Ana-Maria Cretu, Klim Kireev, Amro Abdalla, Wisdom Obinna, Raphael Meier, Sarah Adel Bargal, Elissa M. Redmiles, and Carmela Troncoso. 2025. Evaluating Concept Filtering Defenses against Child Sexual Abuse Material Generation by Text-to-Image Models. *arXiv preprint arXiv:2512.05707* (2025).
- [20] Alejandro Cuevas and Manoel Horta Ribeiro. 2026. Deepfake Pornography is Resilient to Regulatory and Platform Shocks. *arXiv preprint arXiv:2602.02754* (2026).
- [21] Alejandro Cuevas and Manoel Horta Ribeiro. 2026. Deepfake Pornography is Resilient to Regulatory and Platform Shocks. doi:10.48550/arXiv.2602.02754 arXiv:2602.02754 [cs].
- [22] Chi Cui, Yixin Wu, and Yang Zhang. 2026. From Celebrities to Anyone: Characterizing AI Nudification Content, Technology, and Community Dynamics on 4chan. *arXiv preprint arXiv:2606.27234* (2026).
- [23] Alaa Daffalla, Sarah Chao, and Eric Zeng. 2026. Dual-Use AI Face Swap Apps Are Mostly Unsafe: A Systematic Safety Audit. *arXiv preprint arXiv:2605.24735* (2026).
- [24] Mohamed Moustafa Dawoud, Alejandro Cuevas, and Ram Sundara Raman. 2026. From Underground to Mainstream Marketplaces: Measuring AI-Enabled NSFW Deepfakes on Fiverr. In *Proc. of USEC*.
- [25] Periwinkle Doerfler, Andrea Forte, Emiliano De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. 2021. "I’m a Professor, which isn’t usually a dangerous job": Internet-facilitated Harassment and Its Impact on Researchers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–32.

- [26] Nicola Döring, Thuy Dung Le, and Dan J Miller. 2025. Experiences with AI-Generated Pornography: A Quantitative Content Analysis of Reddit Posts. *Archives of Sexual Behavior* (Sept. 2025).
- [27] Brianna Dym and Casey Fiesler. 2020. Social Norm Vulnerability and its Consequences for Privacy and Safety in an Online Community. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 155 (Oct. 2020), 24 pages. doi:10.1145/3415226
- [28] FBI. 2024. Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal. <https://www.ic3.gov/PSA/2024/PSA240329>
- [29] Asher Flynn, Anastasia Powell, Asia A Eaton, and Adrian J Scott. 2025. Sexualized deepfake abuse: Perpetrator and victim perspectives on the motivations and forms of non-consensually created and shared sexualized deepfake imagery. *Journal of Interpersonal Violence* (2025).
- [30] P Brooks Fuller, Kyla P Garrett Wagner, and Farnosh Mazandarani. 2019. Porn wars: Serious value, social harm, and the burdens of modern obscenity doctrine. *Journal of Gender, Social Policy & the Law* (2019).
- [31] Lan Gao, Abani Ahmed, Oscar Chen, Margaux Reyl, Zayna Cheema, Nick Feamster, Chenhao Tan, Kurt Thomas, and Marshini Chetty. 2026. Governance of AI-Generated Content: A Case Study on Social Media Platforms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 1604, 22 pages. doi:10.1145/3772318.3790415
- [32] Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin RB Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno. 2025. Analyzing the {AI} Nudification Application Ecosystem. In *Proc. of USENIX Security Symposium*.
- [33] Se Eun Gong. 2024. South Korea investigates Telegram over alleged sexual deepfakes. NPR. <https://www.npr.org/2024/09/06/nx-s1-5101891/south-korea-deepfake>
- [34] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW2 (2021), 1–35.
- [35] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric. 2025. Characterizing the {MrDeepFakes} Sexual Deepfake Marketplace. In *Proc. of USENIX Security Symposium*.
- [36] Guy Harling, Michael M Chanda, Katrina F Ortblad, Magdalene Mwale, Steven Chongo, Catherine Kanchele, Nyambe Kamungoma, Leah G Barresi, Till Bärnighausen, and Catherine E Oldenburg. 2019. The influence of interviewers on survey responses among female sex workers in Zambia. *BMC medical research methodology* (2019).
- [37] Alexa M Harris, Diego Gómez-Zarà, Leslie A DeChurch, and Noshir S Contractor. 2019. Joining together online: the trajectory of CSCW scholarship on group formation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
- [38] Will Hawkins, Brent Mittelstadt, and Chris Russell. 2025. Deepfakes on Demand: The rise of accessible non-consensual deepfake image generators. In *Proc. of FAccT*.
- [39] Jeff Hearn, Kjerstin Andersson, and Malcolm Cowburn. 2007. Background paper on guidelines for researchers on doing research with perpetrators of sexual violence. *Sexual Violence Research Initiative* (2007).
- [40] Jana Hillert, Lina Sophie Haubrock, Arne Dekker, and Peer Briken. 2024. Web-Based Initiatives to Prevent Sexual Offense Perpetration: A Systematic Review. *Current Psychiatry Reports* 26, 4 (April 2024), 121–133. doi:10.1007/s11920-024-01489-1
- [41] Sylvia Hui. 2026. What to know about UK legal changes aiming to regulate AI-generated nude images. <https://apnews.com/article/uk-grok-regulation-laws-63406274d1c3040bf4da60460577f34c>
- [42] Sexual Violence Research Initiative, Rachel Jewkes, Elizabeth Dartnall, Yandisa Sikweyiya, and others. 2012. *Ethical and safety recommendations for research on perpetration of sexual violence*. Sexual Violence Research Initiative, Medical Research Council.
- [43] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–23. doi:10.1145/3359157
- [44] Julia Kieserman, Cat Mai, Sara Lignell, Lucy Qin, Athanasios Andreou, Damon McCoy, and Rosanna Bellini. 2026. Caught in a Mafia Romance: How Users Explore Intimate Narratives with Chatbots. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

- [45] Valerie A Lapointe, Simon Dubé, Sophia Rukhlyadyev, Tinhinane Kessai, and David Lafortune. 2025. The present and future of adult entertainment: A content analysis of AI-generated pornography websites. *Archives of Sexual Behavior* (2025), 1–19.
- [46] Valerie A Lapointe, Simon Dubé, Aurélie Petit, Tinhinane Kessai, Sophia Rukhlyadyev, Vivianne Gravel, and David Lafortune. 2026. The governance of AI-generated pornography platforms: A content analysis. *New Media & Society* (Feb. 2026), 14614448261421873. doi:10.1177/14614448261421873
- [47] Elizabeth J. Letourneau, William W. Eaton, Judith Bass, Frederick S. Berlin, and Stephen G. Moore. 2014. The Need for a Comprehensive Public Health Approach to Preventing Child Sexual Abuse. *Public Health Reports®* 129, 3 (May 2014), 222–228. doi:10.1177/003335491412900303
- [48] Marc Levy and Barbara Ortutay. 2026. Elon Musk’s xAI sues Minnesota over its first-in-the-nation law banning ‘nudification’ technology. <https://apnews.com/article/minnesota-artificial-intelligence-nudification-x-elon-musk-deepfake-131184be939d540de093b567b12c9e16>.
- [49] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All that’s happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 584–595.
- [50] Aart C Liefbroer and Francesco C Billari. 2010. Bringing norms back in: A theoretical and empirical discussion of their importance for understanding demographic behaviour. *Population, space and place* 16, 4 (2010), 287–305.
- [51] Travis Lloyd, Joseph Reagle, and Mor Naaman. 2025. ‘There Has To Be a Lot That We’re Missing’: Moderating AI-Generated Content on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–24.
- [52] Emanuel Maiberg. 2023. Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale. 404 Media. <https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/>
- [53] Emanuel Maiberg. 2025. AI Dataset for Detecting Nudity Contained Child Sexual Abuse Images. 404 Media (Oct. 2025). <https://www.404media.co/ai-dataset-for-detecting-nudity-contained-child-sexual-abuse-images/>
- [54] Emanuel Maiberg. 2025. Civitai Ban of Real People Content Deals Major Blow to the Nonconsensual AI Porn Ecosystem. 404 Media. <https://www.404media.co/civitai-ban-of-real-people-content-deals-major-blow-to-the-nonconsensual-ai-porn-ecosystem/>
- [55] Emanuel Maiberg. 2025. Civitai, Site Used to Generate AI Porn, Cut Off by Credit Card Processor. 404 Media (online). <https://www.404media.co/civitai-site-used-to-generate-ai-porn-cut-off-by-credit-card-processor/>.
- [56] Emanuel Maiberg and Jason Koebler. 2024. Inside the Booming ‘AI Pimping’ Industry. 404 Media. <https://www.404media.co/inside-the-booming-ai-pimping-industry-3/>
- [57] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (April 2019), 2056305119836778. doi:10.1177/2056305119836778
- [58] Samuel Mayworm, Kendra Albert, and Oliver L Haimson. 2024. Misgendered during moderation: How transgender bodies make visible cisnormative content moderation policies and enforcement in a meta oversight board case. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 301–312.
- [59] Pamela M McMahon. 2000. The public health approach to the prevention of sexual violence. *Sexual Abuse: A Journal of Research and Treatment* 12, 1 (2000), 27–36.
- [60] Bernardo B. P. Medeiros, Malvika Jadhav, Allison Lu, Tadayoshi Kohno, Vincent Bindschaedler, and Kevin R. B. Butler. 2026. Characterizing Resource Sharing Practices on Underground Internet Forum Synthetic Non-Consensual Intimate Image Content Creation Communities. doi:10.48550/ARXIV.2604.12190 Version Number: 1.
- [61] Jaron Mink, Licheng Luo, Natã M Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. 2022. {DeepPhish}: Understanding user trust towards artificially generated profiles in online social networks. In *31st USENIX Security Symposium (USENIX Security 22)*. 1669–1686.
- [62] Jaron Mink, Lucy Qin, and Elissa M. Redmiles. 2026. "Unlimited Realm of Exploration and Experimentation": Methods and Motivations of AI-Generated Sexual Content Creators.
- [63] Jaron Mink, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, and Gang Wang. 2024. It’s Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA)

- (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 778, 20 pages. doi:10.1145/3613904.3641999
- [64] Joe Mullin. 2025. The TAKE IT DOWN Act: A Flawed Attempt to Protect Victims That Will Lead to Censorship. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2025/02/take-it-down-act-flawed-attempt-protect-victims-will-lead-censorship>.
 - [65] National Center for Missing & Exploited Children. 2026. MissingKids.org: Home. <https://www.missingkids.org/home>. Accessed: 2026-01-14.
 - [66] National Center for Missing & Exploited Children. 2026. National Center for Missing & Exploited Children. <https://www.missingkids.org/home?lang=en-US>
 - [67] Olivia B. Newton and Mel Stanfill. 2020. My NSFW video has partial occlusion: deepfakes and the technological production of non-consensual pornography. *Porn Studies* 7, 4 (Oct. 2020), 398–414. doi:10.1080/23268743.2019.1675091
 - [68] Karl-Dieter Opp. 2001. How do norms emerge? An outline of a theory. *Mind & Society* 2, 1 (2001), 101–128.
 - [69] Organization for Transformative Works. 2026. Archive of Our Own. <https://archiveofourown.org/>
 - [70] Elinor Ostrom. 2011. Background on the institutional analysis and development framework. *Policy studies journal* 39, 1 (2011), 7–27.
 - [71] Erin O’Callaghan and Katherine Lorenz. 2024. “I Think I Just like Having Sex”: A Qualitative Study of Sexual Assault Survivors and Their Sexual Pleasure. *Sex Roles* 90, 9 (2024), 1169–1187.
 - [72] Parliament of the United Kingdom. 2025. Data (Use and Access) Act 2025, Section 138. <https://www.legislation.gov.uk/ukpga/2025/18/section/138> Issue: c. 18.
 - [73] Sachin R Pendse, Darren Gergle, Rachel Kornfield, Jonah Meyerhoff, David Mohr, Jina Suh, Annie Wescott, Casey Williams, and Jessica Schleider. 2025. When Testing AI Tests Us: Safeguarding Mental Health on the Digital Frontlines. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1793–1804.
 - [74] Zoë D Peterson, Kaylee P Richmond, Laura Holt, Karen S Beale, and Victoria Cook. 2024. Can a Sex-Positive Sex Education Intervention Reduce Rape-Supportive Attitudes? Formative Research, Pilot Randomized Controlled Trial, and Evaluation of Self-Selection Bias. *American Journal of Sexuality Education* (2024).
 - [75] Aurélie Petit. 2025. The limits of ‘zero tolerance’ policies for animated pornographic media. *Porn Studies* (May 2025), 1–18. doi:10.1080/23268743.2025.2491506
 - [76] Lyne Piché, Jeffrey Mathesius, Patrick Lussier, and Anton Schweighofer. 2018. Preventative services for sexual offenders. *Sexual Abuse* 30, 1 (2018), 63–81.
 - [77] Pieter Haeck. 2026. EU set to ban AI nudification apps in wake of Grok scandal. Politico. <https://www.politico.eu/article/eu-grok-x-elon-musk-ai-nudification-ban-in-wake-of-scandal/>
 - [78] Jeremy Prichard, Richard Wortley, Paul Watters, Caroline Spiranovic, and Joel Scanlan. 2024. The effect of therapeutic and deterrent messages on Internet users attempting to access ‘barely legal’ pornography. *Child Abuse & Neglect* 155 (Sept. 2024), 106955. doi:10.1016/j.chiabu.2024.106955
 - [79] Lucy Qin, Jaron Mink, and Elissa M. Redmiles. 2026. [Additional Materials] “I Thought You Were The Uncensored Place”: Norms, Rules, and Moderation in AI-Generated Sexual Content Communities. https://osf.io/g6a83/overview?view_only=e5acd9ae370c439f857d8ba33b6c51e1
 - [80] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2020. “It’s easier than causing confrontation”: sanctioning strategies to maintain social norms and privacy on social media. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–25.
 - [81] Piera Riccio, Thomas Hofmann, and Nuria Oliver. 2024. Exposed or erased: algorithmic censorship of nudity in art. In *Proc. of CHI*.
 - [82] Joel Scanlan, Lauren C. Hall, Paul A. Watters, Richard Wortley, and Jeremy Prichard. 2026. The Relationship Between a CSAM Warning Messaging Chatbot and User Behavior on Pornhub. *Victims & Offenders* 21, 3 (April 2026), 574–602. doi:10.1080/15564886.2025.2606857
 - [83] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* 26, 10 (2024), 5677–5701.
 - [84] Amartya K Sen. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & public affairs* (1977), 317–344.

- [85] Marko Siitonen. 2007. *Social interaction in online multiplayer communities*. Number 74. University of Jyväskylä.
- [86] Yandisa Sikweyiya and Rachel Jewkes. 2011. Perceptions about safety and risks in gender-based violence research: implications for the ethics review process. *Culture, health & sexuality* (2011).
- [87] Emma Sower, Bronwyn Neal, and Alexis Schmader. 2023. Kink and BDSM Awareness in Sex Offense Treatment. *Journal of Positive Sexuality* (2023).
- [88] Jessica Speed. 2025. Japan struggles with gray zone of AI deepfakes exploiting children. *The Japan Times* (April 2025). <https://www.japantimes.co.jp/news/2025/04/27/japan/crime-legal/deepfake-a-i-child-pornography/>
- [89] Jordan Taylor, Joel Mire, Franchesca Spektor, Alicia DeVrio, Maarten Sap, Haiyi Zhu, and Sarah E Fox. 2025. Un-Straightening Generative AI: How Queer Artists Surface and Challenge Model Normativity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 951–963.
- [90] Jordan Taylor, Joel Mire, Franchesca Spektor, Alicia DeVrio, Maarten Sap, Haiyi Zhu, and Sarah E Fox. 2025. Un-Straightening Generative AI: How Queer Artists Surface and Challenge Model Normativity. In *Proc. of FAccT*.
- [91] Ted Cruz. 2025. S.146 - TAKE IT DOWN Act. <https://www.congress.gov/bill/119th-congress/senate-bill/146>
- [92] David Thiel, Melissa Stroebel, Rebecca Portnoff, Jeffrey Hancock, Cassandra Scyphers, Tim O’Gorman, and Renée DiResta. 2023. *Generative ML and CSAM: Implications and Mitigations*. Technical Report. Stanford Digital Repository. doi:10.25740/JV206YG3793
- [93] Katrin Tiidenberg. 2021. Sex, power and platform governance. *Porn Studies* 8, 4 (2021), 381–393.
- [94] 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 (2023).
- [95] Taro Tsuchiya, Haoxiang Yu, Tina Marjanov, Alice Hutchings, Nicolas Christin, and Alejandro Cuevas. 2026. A Large-Scale Study of Telegram Bots. doi:10.48550/arXiv.2603.24302 arXiv:2603.24302 [cs].
- [96] Rebecca Umbach, Nicola Henry, and Gemma Beard. 2025. Prevalence and Impacts of Image-Based Sexual Abuse Victimization: A Multinational Study. In *Proc. of CHI*.
- [97] 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 *Proc. of CHI*.
- [98] Rebecca Umbach, Nicola Henry, Renee Shelby, Gemma Stevens, and Kwynn Gonzalez-Pons. 2026. AI-generated image-based sexual abuse: Perpetration and consumption across three regions. *Computers in Human Behavior* 179 (June 2026), 108935. doi:10.1016/j.chb.2026.108935
- [99] Rebecca Umbach, Griffin Hunt, John Buckley, Joel Scanlan, Caoilte Ó Ciardha, Ethel Quayle, Ainslie Heasman, Maximilian von Heyden, Elizabeth Letourneau, Donald Findlater, Tegan Insoll, Richard Wortley, Chad Steel, and Abhishek Roy. 2026. Detering Searches for Child Sexual Abuse Material on Google Search and Promoting Help-Seeking. doi:10.48550/ARXIV.2606.06126 Version Number: 2.
- [100] UNESCO. 2018. Comprehensive Sexuality Education to Prevent Gender-Based Violence. UNESCO. <https://www.unesco.org/en/articles/comprehensive-sexuality-education-prevent-gender-based-violence> Last updated April 20, 2023.
- [101] U.S. Supreme Court. 1973. Marvin Miller v. State of California.
- [102] Varun Nagaraj Rao and Dhanaraj Thakur. 2025. *Think Twice Before You Search: Deterrence Messaging Designs to Prevent Searches for Non-Consensual Intimate Images*. Technical Report. Center for Democracy & Technology. <https://cdt.org/insights/think-twice-before-you-search-deterrence-messaging-designs-to-prevent-searches-for-non-consensual-intimate-images>
- [103] Miranda Wei, Christina Yeung, Franziska Roesner, and Tadayoshi Kohno. 2025. "We’re utterly ill-prepared to deal with something like this": Teachers’ Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery. In *Proc. of CHI*.
- [104] DJ Williams, Emily Prior, and Jenna Wegner. 2013. Resolving social problems associated with sexuality: Can a “sex-positive” approach help? *Social work* 58, 3 (2013), 273–276.
- [105] DJ Williams, Jeremy N Thomas, and Emily E Prior. 2015. Moving full-speed ahead in the wrong direction? A critical examination of US sex-offender policy from a positive sexuality model. *Critical Criminology* (2015).

- [106] Matthew Williams. 2006. *Virtually criminal: Crime, deviance and regulation online*. Routledge.
- [107] Jason Wilson. 2026. Hundreds of nonconsensual AI images being created by Grok on X, data shows. *The Guardian* (Jan. 2026).
<https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images>
- [108] Sandra R Wilson, Nancy L Brown, Carolina Mejia, and Philip W Lavori. 2002. Effects of interviewer characteristics on reported sexual behavior of California Latino couples. *Hispanic Journal of Behavioral Sciences* 24, 1 (2002), 38–62.
- [109] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
- [110] Hyunsu Yim. 2024. South Korea to criminalise watching or possessing sexually explicit deepfakes. Published: Reuters.
- [111] Jiayi Zeng, Mengmeng Wu, and Xingyu Lan. 2026. The Hidden Desires: Exploring Chinese Women’s Erotic Role-Play with AI Chatbots. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, Barcelona Spain, 1–15. doi:10.1145/3772318.3791580
- [112] Andrew Zolides. 2021. Gender moderation and moderating gender: Sexual content policies in Twitch’s community guidelines. *New Media & Society* 23, 10 (2021), 2999–3015.

A Methodology (Continued)

Interested participants were provided a screener survey in which they answered questions on (1) their use of AI for generating sexual content, (2) their demographic background. From this pool of submissions, participants were invited to schedule a 90-minute interview with the research team over an end-to-end encrypted call. When reviewing responses to select participants, we prioritized diverse use cases for AIG-SC creation. Overall, we collected 30 hours and 36 minutes of interview data; interviews lasted 76 minutes on average. In most cases, participants were paired with an interviewer with a matching gender presentation [36, 108].

Analysis. We conducted reflexive thematic analysis [6] and followed an inductive coding process. This began with each author reviewing a selection of interview transcriptions to explore themes within the data. Through this process, each author created an independent preliminary codebook. They then met, discussed themes that emerged, and merged their codebooks into a consolidated version that was used to code an initial set of three interview transcriptions that were triple-coded by all three authors. Afterwards, the authors met to discuss their coded transcriptions and resolved any discrepancies that emerged. This also resulted in minor revisions to the codebook. The authors then coded two more interview transcriptions using the codebook. They then met to further resolve discrepancies and refine the codebook. This discussion produced a finalized codebook (see [79]). The three authors then divided the remaining transcripts and independently coded them. While we report on commonalities in what we observed, we emphasize that each participant’s values and boundaries are deeply personal and vary widely. We encountered a range of perspectives on AIG-SC and advise against treating AIG-SC creators as a monolith.

A.1 Demographics

Geographic Location. At the time of their interview, the majority of our participants resided in North America: the United States ($n=9$), Canada ($n=3$), and Mexico ($n=2$). They also resided in the following countries: Australia, Brazil, Chile, China, Colombia, France, India, Indonesia, The Netherlands, the United Kingdom ($n=1$ each).

Age. Two participants were in the 18-21 age group, four participants were in the 22-24 age group, twelve participants were in the 25-34 year old age group, five participants were in the 35-44 age group, and 1 individual was in the 45-54 age group.

Optional Responses. Participants were provided open-text response fields for questions on race & ethnicity, gender, and their sexual orientation.

Slightly over half ($n=13$) of our participants self-described as White or Caucasian, including those who self-described as “White & Hispanic”, “Mexican caucasian”, and “presenting as white”. Six responded that they are Hispanic and/or Latin American and five responded that they are Asian or South Asian.

Table 1 displays participants’ responses to optional questions on gender and sexual orientation. In order to present summary statistics, we categorized responses on gender and sexual orientation based on their closest categorical approximation. However, these results do not fully capture the richness of responses provided in self-descriptions.

Lastly, participants could optionally respond if they had a disability (including physical, mental health, neurodivergence). While three preferred not to respond, eleven participants responded “no” and ten responded “yes”.

How would you describe your sexual orientation?	# Participants
Heterosexual	16
Bisexual/Gay/Lesbian/Pansexual/Queer	7
No Response	1
How would you describe your gender?	#
Man	18
Non-Binary	4
Woman	3
No Response	1

Table 1: Gender and sexual orientation. * Several participants identify as multiple genders, and thus, these counts are not mutually exclusive.

Category	Count
Gig work	8
Science and technology	7
Other	6
Arts, culture and entertainment	5
Communications	4
Sales	3
Health and medicine	3
Architecture and engineering	2
Education	2
Business, management and administration	2
Installation, repair and maintenance	2

Table 2: Occupational background (participants could select more than 1 category).

Education Level	Count
Did not finish high school	1
High school graduate (diploma or equivalent)	4
Some college or Associate degree (e.g., A.A., A.S.)	7
Trade, technical, or vocational training	2
Bachelor’s degree (e.g., B.A., B.S., B.Eng.)	7
Postgraduate degree (e.g., Master’s, J.D., Ph.D.)	3

Table 3: Educational background.

B Additional Results

B.1 Comparing Text-Based Communities to Visual-Based Communities

In contrast, text-focused communities were much less likely to share content. As text-content within AIG-SC communities often consisted of highly personalized role-playing scenarios [44, 62, 111], participants reported that role-playing interactions were often “*a reflection of who you are, what you like, what you don’t like, what you want to explore... this specific collection and configuration is only going to have full complete relevance to you*” (C04). In some cases, it could even be “*considered a little bit rude to just like share your responses*” (C09). Instead, participants reported that it’s much more common to share either jailbreaks as they are often required for AIG-SC text workflows, or ways to customize roleplaying scenarios, such as character cards (see Section 5).

Additionally, unlike visual-focused communities, text-focused communities held reservations about the effect of AI-generated erotic fiction on human writers: “*fic sharing sites like Ao3⁶ are getting flooded with generative AI fanfiction... flooding what is meant to be a human-creativity sharing space with low-effort content*” (C06). Perhaps because members of text-focused communities were already writers or involved in communities in which writers depended on patronage, they were often aware of AIG-SC’s potential to displace human-created content, and intentionally refrained from sharing AIG-SC. For example, C05 explained that they “*try to keep the story for myself*” to avoid competing with writers.

B.2 Community Rules

Table 4 provides a summary of the rules we encountered that are discussed in Section 6.2. We discuss additional rules below.

Age Attestation. *Community A* and *Community B*, which focused on visual content creation, also had public rules that by participating in the community, users self-attest to being at least 18 years of age. Both communities included warnings that users may be banned if they are discovered to be minors.

Obscenity. In both communities that focused on visual content, there were rules against obscenity such as “*rules against gore and violence*” (PB06). *Community A* did not include a definition of what qualifies as obscenity but referenced U.S. obscenity laws [101], while *Community B* had rules against the depiction of bestiality, sexual violence, and excessive violence.

⁶Archive of Our Own (Ao3) is a platform for sharing user-generated stories [69].

Community Rules	A	B	C
AIG-SC			
AI content sharing only	●		
No use of real likenesses	●	●	
No content that depicts minors	●	●	●
No obscenity	●	●	
Members must follow Discord's TOS	●	●	
Members must follow the TOS/License of AI models	●	●	
Assets			
No minor enabling tools			●
No real people	●		
Peer			
Duty to report violations		●	●
No kink shaming		●	●
No discrimination		●	●
No harassment	●	●	●
No political/moral debates		●	●
User			
Server not liable for posts	●		
Age +18 Self-Attestation	●	●	

Table 4: Community Rules — We present communities’ posted rules. ● notes that such a rule is present and comprehensive, whereas ● notes that the rule is partially covered, or covered to a lesser extent (e.g., while Community A, B, and C all have rules against content depicting minors, only B also has rules against subjects appears in contexts, like classrooms, that could imply the subject is a minor).

Peer Conduct & User Requirements. Several community rules address appropriate conduct. In particular, several rules instantiate *Norm #3* and explicitly ban kink shaming (*Community B*, *Community C*), discrimination by race, sex, sexual preference, and other aspects of sexuality (*Community B*, *Community C*), and general forms of harassment (*Community A*, *Community B*, *Community C*). In terms of specific behaviors, only *Community C* explicitly requires members to ask for consent before sending a direct message, perhaps to avoid instances of abuse where creators are approached and solicited to create AIG-NCC: “And so they’ll like DM that person and they’ll be like, ‘Hey, I’ll give you like forty bucks if you give me instructions on how to make a model based off someone’” (B01).

Several communities also ban debates around politics and morality (*Community B*, *Community C*). Moderators have noted that this is to prevent instances of perceived political extremism such as C01 trying to prevent “[alt]-right and MAGA people coming to spew their hate”.

However, given that C14 describes recent takedowns of AIG-NCC among offending websites as “all this sort of, you know, payment processor and, you know, political whatever”, these rules may also silence discussion of anti-abusive personal commitments through which anti-abusive norms could be established within the broader community.