Real Time Threats

Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms

Robert Gorwa
Dhanaraj Thakur

November 2024

ROBERT GORWA

Research Fellow at the WZB Berlin Social Science Center and Non-Resident Fellow at the Center for Democracy & Technology.

DHANARAJ THAKUR

Research Director at the Center for Democracy & Technology.

Real Time Threats

Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms

Robert Gorwa and Dhanaraj Thakur*

Safe Online

Contents

Executive Summary

I n recent years, a range of new online services have emerged that facilitate the 'livestreaming' of real-time video and audio. Through these tools, users and content creators around the world can easily broadcast their activities to potentially large global audiences, facilitating participatory and generative forms of collaborative 'live' gaming, music making, discussion, and other interaction. The rise of these platforms, however, has not been seamless: these same tools are used to disseminate socially problematic and/or illegal content, from promotion of self-harm and violent extremism to child sexual exploitation and abuse (CSEA) materials.

This report examines the range of trust and safety tools and practices that platforms and third-party vendors are developing and deploying to safeguard livestreaming services, with a special focus on CSEA prevention. Moderating real-time media is inherently technically difficult for firms seeking to intervene responsibly: much livestreaming content is "new", produced on the spot, and thus by definition not "known" and possible to match against previously identified harmful material through hash-based techniques. Firms seeking to analyze livestreams instead must do so with comparatively inefficient and potentially flawed predictive computer vision models, working creatively with the stream audio (e.g., through transcription and text classification), and/or through other emerging techniques, such as "signals"-oriented interventions based on the behavioral characteristics of suspicious user accounts.

Based on a review of publicly available documents of livestreaming platforms and vendors that offer content analysis services, as well as interviews with persons working on this problem in industry, civil society, and academia, we find that industry is taking three main approaches to address CSEA in livestreaming:

- **Design based approaches —** Steps taken before a user is able to stream, such as implementing friction and verification measures intended to make it more difficult for users, or suspicious users, to go live. For example, some platforms require a user to have a threshold number of followers or subscribers before they can livestream to prevent an actor from spontaneously creating an account and livestreaming harmful content.
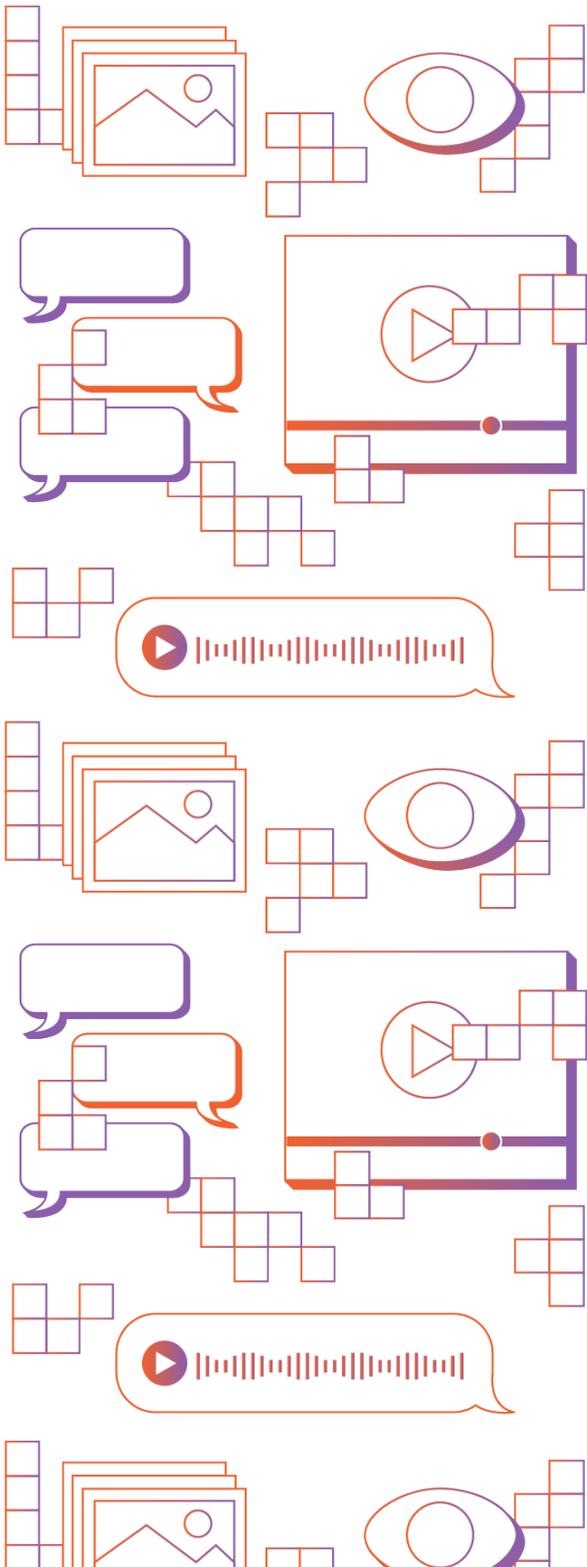
- **Content analysis approaches** — Various forms of manual or automated content detection and analysis that can work on video, audio, and text as content is livestreamed. Examples include taking sample frames from livestreams and seeing if they match hashes of known CSEA material; using machine learning classifiers to detect CSAM on live video; and employing predictive analysis of text transcriptions of live audio or user chats in livestreams.

- **Signal based approaches** — Interventions based on the behavioral characteristics and metadata of user accounts. For example, platforms may share certain account metadata to help identify bad actors as they move from platform to platform or use signals to identify accounts engaged in potentially suspicious behavior that prompts further investigation.

In part because of the challenges of livestream content detection, the way in which industry tackles the problem of CSEA and other harmful content is evolving. As one interviewee put it, the idea is for firms to engage more actively in reducing the ability to use their platform for CSEA dissemination, not only engaging in a detect and report mode but also, aspirationally, towards a predict and disrupt model of trust and safety more akin to that used in areas such as cybersecurity and fraud.

Industry approaches to CSEA raise several concerns. First, there is a general trend to eschew transparency and clarity in how these systems operate and are deployed, ostensibly to prevent bad actors from circumventing them, but potentially to the detriment of victims, users, policymakers, and other stakeholders. Second, and related to the first point, it is almost impossible to determine how effective these approaches are, what gaps they leave, whether they result in overmoderation of legitimate content, and how well they serve the needs of all stakeholders. Third, these approaches introduce significant security, privacy, free speech, and other human rights risks that can undermine the safety of the minors that they are meant to protect as well as that of users in general.

To help address these concerns, we highlight four areas for improvements:

1. **Greater transparency is needed to help evaluate and improve efforts to address CSEA on livestreaming platforms.** For example, there are currently no performance metrics that firms can use to test and compare the accuracy of the measures they take or that experts, policymakers, and researchers can use to gain a better understanding of their efficacy, as well as the extent of what is really possible.

2. **Vendors and livestreaming platforms should be explicit about the limitations of automated approaches to detecting and addressing CSEA.** In so doing, platforms can improve their trust and safety systems by ensuring human reviewers are appropriately involved and allowing them to make nuanced decisions based upon context and other information.

3. **Focus on design interventions that empower users including minors.** The needs of streamers to protect themselves from being targeted with or being used to distribute CSEA are worthy of greater attention when it comes to design based solutions. For example, one design based approach that was not raised in our discussions with industry is to provide users, particularly minors, with the right set of tools and reporting mechanisms to help them protect themselves and others.

4. **Multistakeholder governance models can improve accountability of approaches to address CSEA on livestreaming.** Best practice frameworks around the implementation of these systems could be developed not only through the continuing work of organizations like the Tech Coalition, but also through critical multistakeholder engagement in fora that not only involve child safety organizations, but also organizations actively engaged on a broader set of digital rights and civil liberties.

Addressing the problem of CSEA in general and on livestreaming platforms is critically important given the impacts on children, parents, and their communities, so this is a hugely consequential and high-stakes area of platform governance. Vendors and industry alike are understandably eager to show that they are developing innovative new tools to address CSEA and other harmful content, but poor implementation (or poor design, with systems that are fundamentally flawed) will decrease, rather than increase, policymaker and public confidence in platforms' trust and safety over the longer term. Better understanding of the measures platforms are taking on livestreaming platforms, along with increased multistakeholder engagement, will improve trust and safety systems in ways that minimize the risk of CSEA in livestreamed content, while also minimizing unintended impacts on ordinary users.
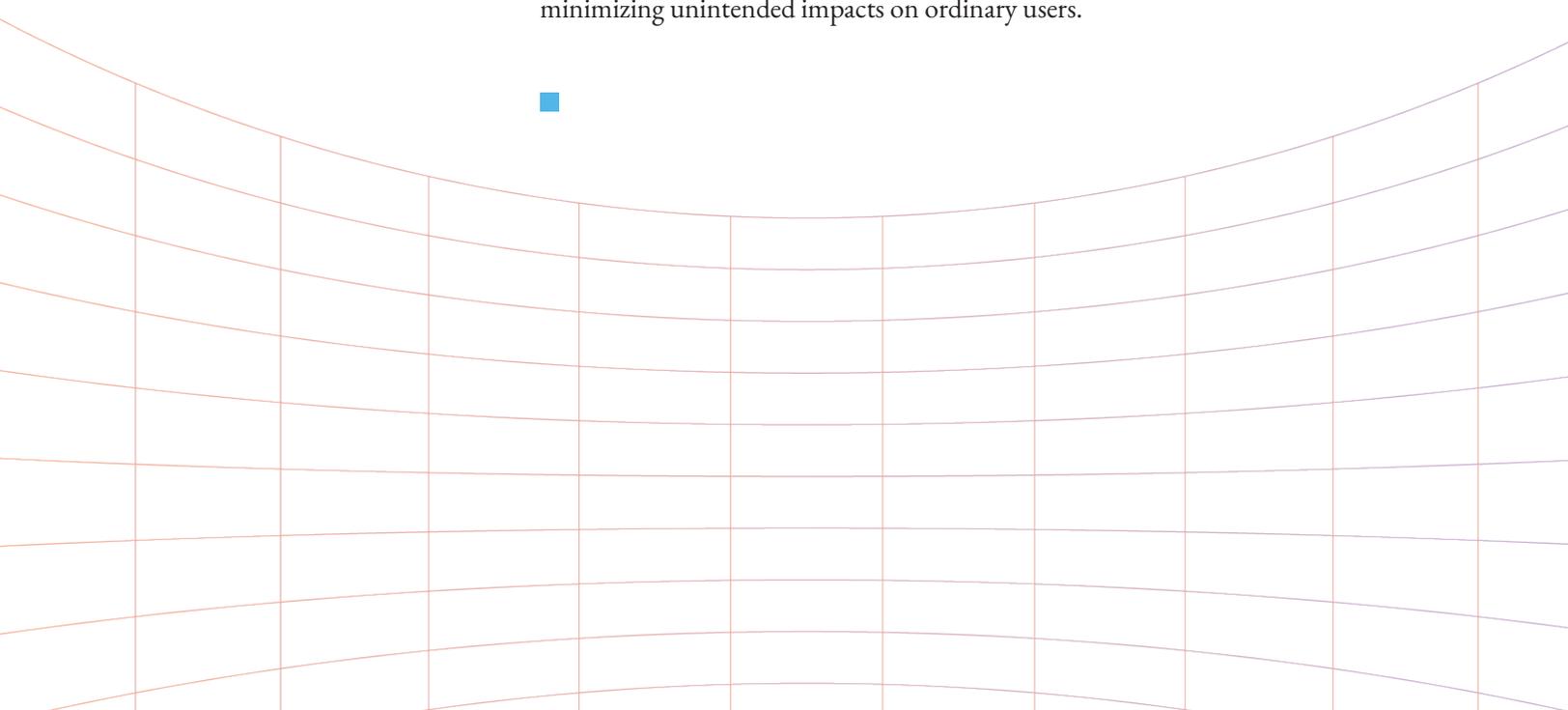
Introduction

I n recent years, a range of new online services have emerged that permit the real time transmission of video and audio. Through these livestreaming tools, users and content creators can distribute content as it is created (rather than finalizing content before beginning distribution). Streamers can broadcast their activities in real time to large global audiences, facilitating participatory and generative forms of collaborative "live" gaming, music-making, discussion, and other interaction. This is typically done from one to many persons in various ways, such as public streams or in private groups. Most livestreaming services use a combination of compression, encoding, and content delivery networks to distribute content around the world (Cloudflare, n.d.).

As livestreaming services become more prevalent and popular, especially with younger users, stakeholders in government, civil society, and industry have directed a growing amount of attention to the "trust and safety" practices of the companies operating them. In particular, law enforcement agencies, civil society groups, and others have consistently expressed concern about the possibility of livestreaming platforms being used to facilitate various forms of child sexual exploitation and abuse (CSEA) (Setter et al., 2021).

CSEA is generally understood to include the production, dissemination, and possession of child sexual abuse material/images (CSAM, or content containing sexually explicit activities with children); online grooming of children for sexual purposes; sextortion; and online prostitution involving children (Quayle, 2020). Livestreaming of CSEA therefore involves the "real-time producing, broadcasting, and viewing of child sexual abuse and is related to sexual exploitation through prostitution, sexual performances, and producing CSEA." (Drejer, Riegler, et al., 2024). While comprehensive empirical data is scant, there is evidence that various types of livestreaming platforms have become a vector for the production, dissemination, and consumption of CSEA (Insoll et al., 2021).

In the context of non-livestreamed, stored content, internet intermediaries have in the past decade implemented a set of minimum best practices intended to counter the proliferation of child sexual abuse imagery online and to comply with formal legal regimes in countries like the United States. Some of that work has occurred through global multistakeholder collaboration facilitated by organizations like
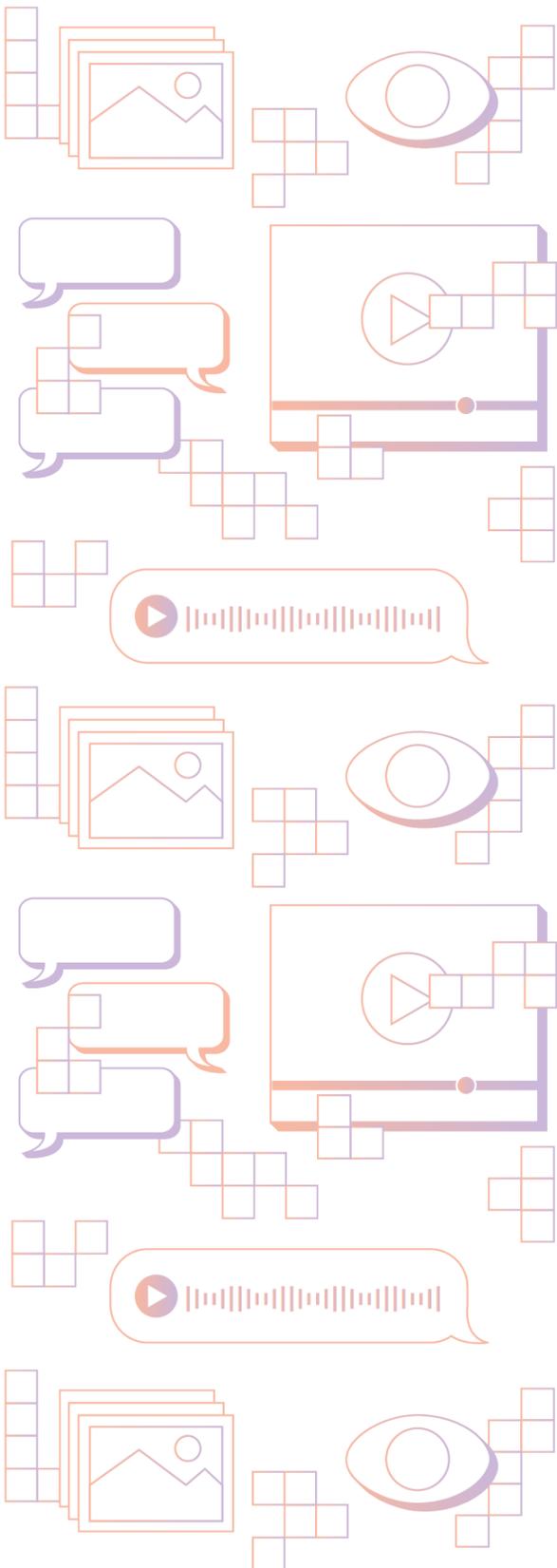
the WeProtect Global Alliance and the National Center for Missing and Exploited Children (NCMEC). At the core of these developments are machine learning techniques that facilitate the computational matching or "fingerprinting" of confirmed child abuse images. Matching is one approach for which machine learning models are used to analyze content online. It allows companies to compare the fingerprints of suspected content to "hash databases" of actual CSAM that are continually maintained and updated by government-linked civil society organizations in various jurisdictions. In other words, matching allows companies and others to identify a piece of content as being identical or sufficiently similar to CSAM that already is known (Shenkman et al., 2021).

Livestreaming, however, poses inherent technical challenges for firms seeking to intervene responsibly against child sexual abuse material. Much livestreaming content is "new", produced on the spot and thus by definition not "known" and possible to match against existing material through hash-based techniques. This fundamental problem removes the most reliable tool for automated content analysis from industry's trust and safety tool boxes (Farid, 2022; Gorwa et al., 2020). Firms seeking to analyze live video must instead often do so using the second main application of machine learning to content analysis — predictive models. These models recognize the characteristics or features of a piece of content based on the machine's prior learning. However, even when applied to static images, such predictive computer vision models or related techniques are often flawed (Shenkman et al., 2021). Those flaws are only magnified in the context of live video. In addition, applying these tools at scale to tens of thousands or even more simultaneous streams is computationally difficult even for well resourced actors, potentially introducing significant latency and quality issues that could undermine the entire value proposition of live-oriented products.

This report explores how the industry is responding to these challenges and some of the privacy and other implications of the solutions they have adopted. The Center for Democracy & Technology (CDT) has closely followed the use of automated content analysis tools, both analyzing their potential value and their implications for human rights and freedom of expression (Duarte et al., 2017; Shenkman et al., 2021). Here, we explore the range of trust and safety tools and practices that are being developed by various platform actors in order to safeguard their livestreaming offerings, seeking to provide an initial mapping of a little-examined and generally opaque ecosystem that nonetheless could have significant impacts on the online experience and rights of many internet users worldwide. In addition to the efforts of major livestreaming platforms, there has been a proliferation of 'safety tech' vendors and other third-party groups pushing a wide range of technical solutions to detect and address CSEA, which we also consider in this report.

We explore the range of trust and safety tools and practices that are being developed by various platform actors in order to safeguard their livestreaming offerings, seeking to provide an initial mapping of a little-examined and generally opaque ecosystem that nonetheless could have significant impacts on the online experience and rights of many internet users worldwide.

After a brief introductory survey of existing research, the second part of this report outlines the automated content analysis techniques livestreaming services currently use to detect child sexual and exploitation abuse. These efforts include (1) design interventions intended to create friction in the process of livestreaming content and thereby reduce the potential for abuse, (2) various techniques for stream audio and video analysis, and (3) metadata and behavioral profiling measures oriented around user accounts that increasingly shift away from detecting problematic content and toward a threat intelligence approach. In the third section, we provide a policy-oriented discussion of these techniques, including an assessment of the most important emerging trends in this ecosystem. These include the increasing use of user profiling, metadata, and "behavioral signals" by platforms to try to identify and intervene against bad actors. We discuss prospective transparency, bias, privacy, and human rights concerns; data quality issues affecting the usage of these systems; and some ways in which platforms are — and could better — implement oversight of their emerging efforts to minimize unintended impacts on ordinary users.

I. Research on Livestreaming and Child Sexual Exploitation and Abuse
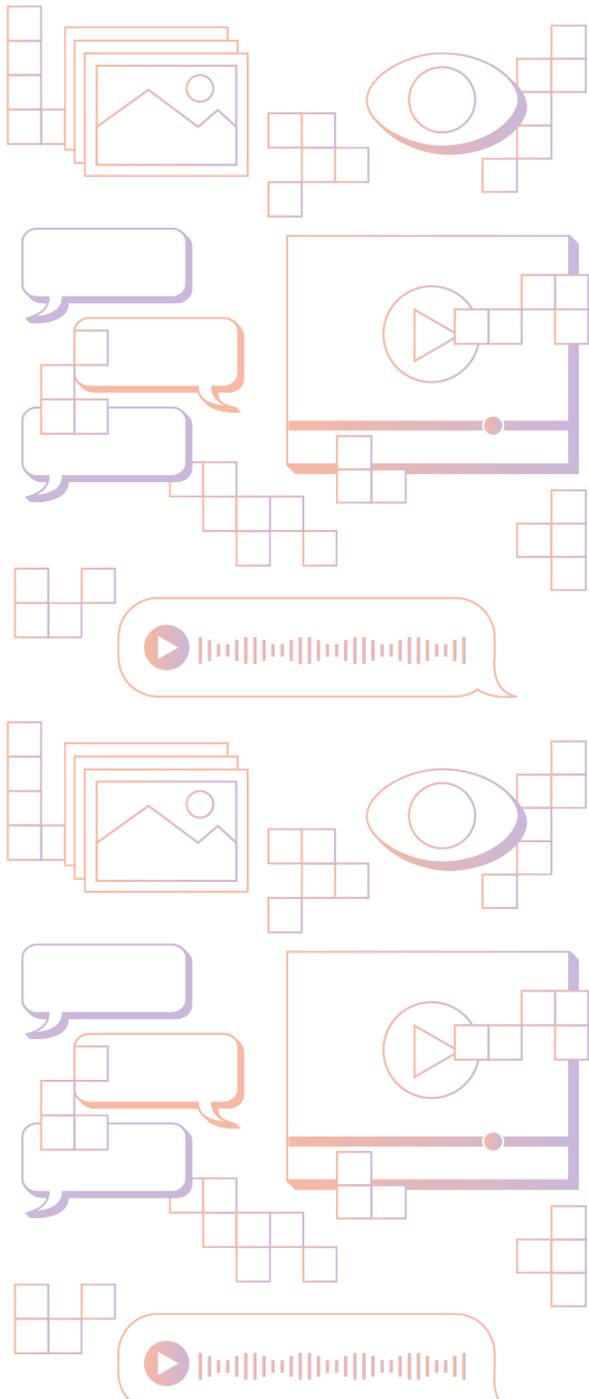
The research on child sexual abuse and livestreaming is difficult to navigate as it often collapses different types of sexual content depicting minors, as well as a wide range of online services with vastly different affordances and underlying technical architectures. While the topic has received ample international media attention, the academic literature on livestreaming and child abuse remains very small and underdeveloped: a 2024 review paper found only 8 peer-reviewed articles on the topic (Drejer, Riegler, et al., 2024).

Even these articles are generally oriented around a specific modality of transnational, online-mediated sex trafficking, with an overwhelming focus on the Philippines (Drejer, Riegler, et al., 2024). The overarching scenario here is that children in low-income contexts enter into various forms of underage camming and sex work, potentially pressed by family members or otherwise seeking to survive in a context of desperate poverty with unclear dynamics of coercion (Christensen & Woods, 2024). Ample non peer-reviewed research published in collaboration with law enforcement has described some of the contours of this phenomenon, where live "viewings" mediated via video-calling platforms are sold to international audiences (Teunissen et al., 2021; Teunissen & Napier, 2023). A review of the 30 public cases prosecuted for this kind of online sexual abuse in the United Kingdom from 2013-2022 shows that coordination for these transactions can occur through a multitude of channels, including online dating forums, adult content sites, and peer-to-peer messaging apps (Celiksoy et al., 2023). Europol press materials have referred to this kind of activity as "live distant child abuse", and this is the central focus of a number of non-governmental organizations' work on international sex trafficking and child sexual abuse and exploitation online (Europol, 2024).

A second, less developed modality of child sexual abuse's intersection with real-time video sharing is less about direct coercion, and more about indirect manipulation and trickery. This work is characterized by a focus on the emerging notion of online "grooming" (Salter & Sokolov, 2024). The concept is increasingly used in the "live" context to describe activity where young people active in various online spaces (e.g., multiplayer online games, online forums) are approached by other users with the goal of building up a relationship with them, then eventually convincing or coercing them to join video calls where they are pressured into sexual activity (Drejer, Sabet, et al., 2024). This content is considered "self-generated" in that it is typically produced by young persons themselves, and much coverage of this kind of "grooming" focuses on "near" high-income contexts rather than the low income international countries at the center of the "live distant child abuse" conversation (Vallance, 2024).

The dynamics of self-generated sexual content production by minors can be ambivalent, however, especially when it comes to older youth and their important and legitimate efforts to pursue their sexual autonomy online (Quayle, 2022). Complicating matters further, multiple complex social and economic factors motivate the creation of various forms of self-generated underage sexual content (Cooper et al., 2016). As some recent research by internet policy and security experts has highlighted, major user-generated content platforms like Instagram have recently become home to underage users who sell or exchange sexual material depicting themselves (Thiel et al., 2023).

Robust and independent cross-platform incidence data about the prevalence of this kind of "self generated" underage sexual material in the livestreaming context does not currently exist. Substantial coverage of livestreaming and child sexual abuse issues in the international media in recent years, however, has highlighted some of the potential dynamics at play, including on a range of popular public-facing platform services frequently used by youth in the United States and beyond. One high-profile investigation by Bloomberg in 2022 investigated the way in which young Twitch streamers were being harassed, tracked, and in some cases, compelled, convinced, or tricked into sexual acts (D'Anastasio, 2022). Reports of "sextortion" have been linked to certain Discord servers (Boburg et al., 2024; Goggin, 2023). TikTok's own internal investigation suggested that children were stripping on its TikTok Live service in exchange for online gifts (Allyn et al., 2024).

Journalists have also documented how the affordances of certain platforms have facilitated the wider dissemination of child sexual abuse material. Young streamers may believe that streams are completely ephemeral, but some platforms have built-in features that allow stream snippets to be saved and made discoverable after the fact. For a 2024 Bloomberg investigation, reporters partnered with the Canadian Centre for Child Protection to investigate Twitch's "Clips" archive, short videos of up to 20 seconds that had been made by viewers of their favorite stream moments, and found that a surprisingly large number (7.5% of a sample of 1100 "clips") could be classified as featuring sexually explicit depictions of minors (D'Anastasio, 2024; Winslow, 2024). Reporting has relatedly suggested that at least some livestreaming contexts may involve the playback of previously obtained sexual material depicting minors in live video and audio forums. An NBC investigation into Discord suggested a complex ecosystem of bad actors that included "'hunters' who located young girls and invited them into a Discord server, 'talkers' who were responsible for chatting with the girls and enticing them, and 'loopers' who streamed previously recorded sexual content and posed as minors to encourage the real children to engage in sexual activity" (Goggin, 2023).

Some reporting and research looking at livestreaming has also addressed real-time networked forms of harassment and trolling, where malicious users flood into new channels and seek to cause trouble for certain streamers (Han et al., 2023), especially women and those from racialized communities (Jackson, 2019; Ruberg, 2021). This harassment can potentially involve the sharing of child sexual abuse material, either to inflict harm on potential viewers or to "nuke" certain streams and get them shut down by moderators or platform's automated content detection. Journalists at 404 Media have shown how this strategy has recently been used by various hacking and fraud groups to shut down rival Discord servers (Cox, 2024a).

Report Scope and Methods

Overall, in this small yet nonetheless growing body of research and journalistic investigation, the term "livestreaming" is used to refer to at least seven related yet distinct categories of platforms:

- Major, general purpose social networks that have livestreaming products or "surfaces" intended for public or large-audience broadcasting (e.g., TikTok Live, Instagram Live, Facebook Live).

- Specific live video sharing platforms which started out largely in the online gaming context but now host a wide-range of entertainment and commentary (e.g., Twitch, Kick, Discord).

- Platforms for audio or video streaming that are built around the idea of an internationally broadcast live event like a talk or a concert (e.g., Clubhouse, Spotify Live).

- General purpose video calling/video conferencing platforms (e.g., Zoom, Teams, Skype, Jitsi, Webex).

- Peer-to-peer direct messaging platforms that also have direct or small-group live calling functionality (e.g., Facetime Video, WhatsApp, Signal, Telegram).

- "Random video chat" applications in the tradition of ChatRoulette that match up two random users for a video call, usually as web browser applications (e.g., Shagle, ChatRandom, ChatHub and other services taking up the Omegle/Chatroulette torch).

- Websites or apps specifically tailored to "in real time" adult sexual performance and pornography (e.g., StripChat, Chaturbate).

The extant research also suggests various related yet distinct modalities of underage sexual content production and dissemination as it pertains to these different services. These include:

- "Live distant child abuse" or real-time video enabled sex trafficking (where the research is often concerned with viewers of livestreams in high-income contexts with facilitators in low-income jurisdictions).

- Non-live child sexual abuse material (videos or images) "re-broadcast" through livestreaming services or distributed via supplementary livestream surfaces (e.g., via links in a chat that accompanies a livestream).

- Self-generated underage sexual content (where the main concern of researchers are creators in high-income jurisdictions).
   - Created and disseminated via livestream (private or public facing).
   - Disseminated in non-live setting after social interactions on a livestreaming service (private or public facing).

The seven different types of platforms are sufficiently different in their core functionality and technical architectures that no single research project can comprehensively speak to them all. Complicating matters, data on the actual prevalence of the different sub-categories of activity remains scarce and of limited quality. Non-governmental organizations have made some efforts to measure the prevalence of CSEA including the use of livestreaming in places like the Philippines, with a 2022 IJM survey of a representative sample of 3600 households suggesting that 1% of the Philippines' under-18 population could be involved in the industry. (International Justice Mission & University of Nottingham Rights Lab, 2023). A coordinated European law enforcement operation in summer 2024 worked with 12 years of data on "criminal networks sexually exploiting children in the Philippines", analyzing information on approximately 12,000 accounts that could be linked to 197 individuals in the European Economic Area, the UK, and the US (Europol, 2024). The transparency reports of popular livestreaming services do not distinguish between self-generated underage sexual content and other forms of CSAM being shared on their services. For instance, Twitch's 2023 report states that the firm took action in 12,801 cases in the first half of 2023 under their "Youth Safety Policy", which includes "illegal CSEA material as well as content and material that is not illegal but violates our Community Guidelines by endangering minors" (Twitch, 2023). It is not clear if these cases involve pieces of content, accounts removed, or both.

Our goal in this report is not to assess the prevalence of the CSEA in livestreaming but to examine how companies and others address the problem. To do that, we first conducted a review of publicly available documentation released by major platforms in the seven categories discussed earlier, particularly those that have functionality for public-facing "broadcast" livestreaming products. We also reviewed the practices of

porn sites, private calling providers, or other services, and integrated insights pertaining to their trust and safety efforts of these companies where possible.

This involved analysis of technical reports, white papers, press releases, blog posts, and other types of industry material put online by firms, as well as by emerging associations like the Tech Coalition and third-party "safety tech" vendors that work with both governments and/or different platforms. We also conducted 15 interviews with industry leaders, vendors, and civil society experts, asking these individuals to reflect on their current tools and practices (where applicable), as well as the challenges that they face in their day-to-day efforts to safeguard streams. Insights from these discussions were refined in a half-day multistakeholder workshop hosted by CDT in June 2024.

As noted earlier, our analysis focuses on CSEA, which includes child sexual abuse material/imagery (CSAM). The latter is subject to a well established governance and legal regime. However, CSEA also includes problems such as grooming. Researchers, for example in critical legal and sexuality studies, distinguish between "conduct" versus "contact" (Baines, 2019; McAlinden, 2006), which alludes to activity like grooming that isn't inherently violent or illegal but could potentially lead to future abuse or violence under certain conditions. Our research showed that firms are, at least to certain degrees, increasingly deploying trust and safety measures seeking to counter this kind of potentially unsafe contact, and we thus have included these practices in the scope of this research where possible as well.
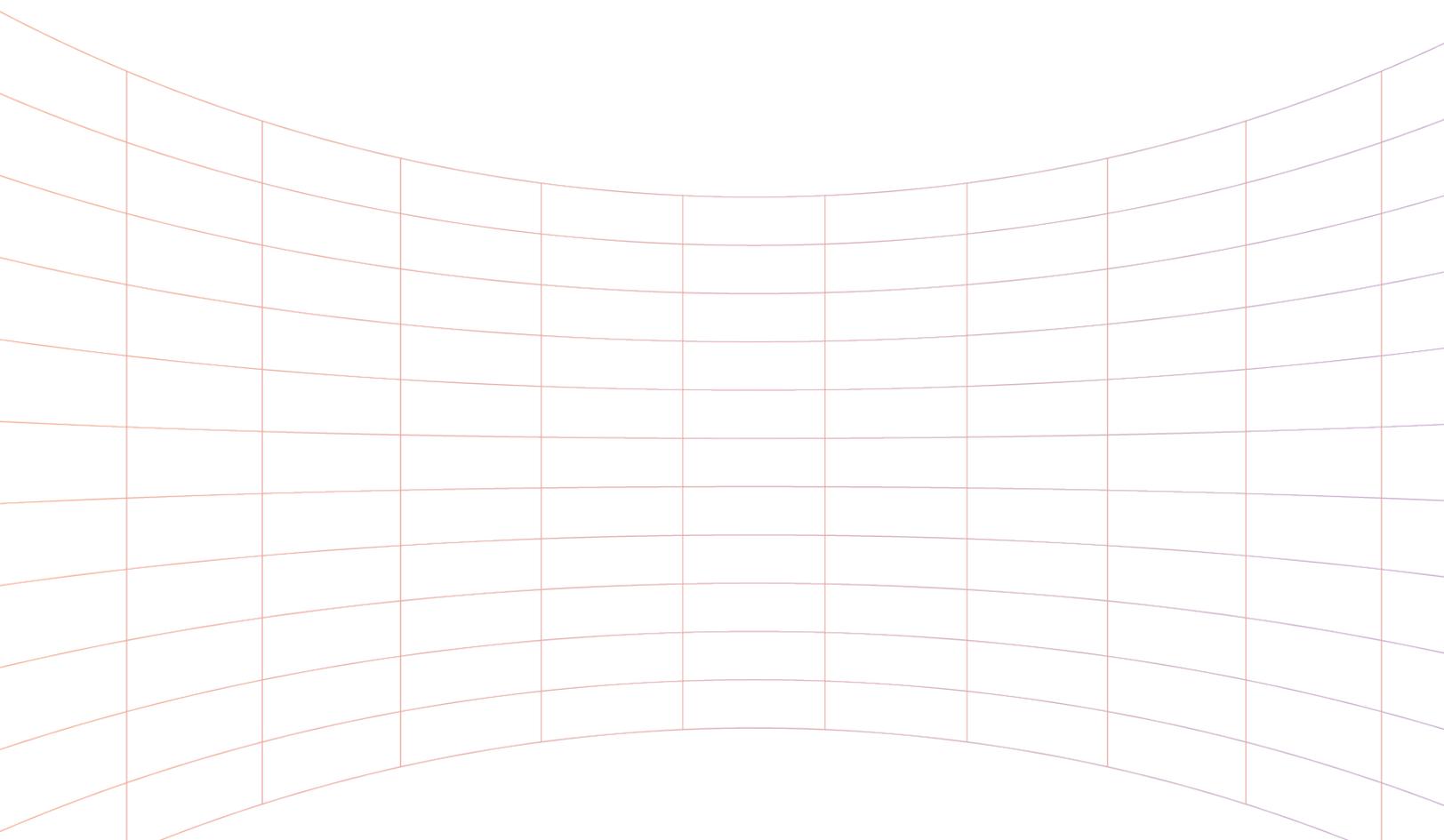
II. Safeguarding Streams: A Trust and Safety Overview

Companies operating services in all of the seven aforementioned "livestreaming" categories have actively been deploying various tools and practices to enforce their policies prohibiting sexual content depicting minors. In general, companies with livestreaming services prohibit CSEA and CSAM on their platforms. This is often explained in the platform's terms of service or community guidelines with prohibitions against sexual content related to young people including content depicting specific actions or behaviors. Some platforms (e.g., Meta) offer specific guidance against content containing CSEA and CSAM (Meta, 2024a). The platforms' terms of services also typically explain that the companies reserve the right to take actions against such content and/or the accounts that disseminate it. These activities are part of their trust and safety operations (or T&S), a term increasingly used by technology companies to refer to the rule-making, rule-enforcement, and system design they deploy to prevent and/or police types of user behavior that go against their policies or local laws (Caplan, 2023; Denyer Willis, 2023).

Academic work on content governance on certain popular livestreaming platforms like Twitch has pinpointed some of the key affordances that make the effective implementation and enforcement of content policies in real-time video environments potentially more difficult when compared to other platform products. One major characteristic of the live context is *ephemerality*: unless the streamer or someone in the audience is using a (usually third-party) tool to record the stream in some way, activity that happens in the stream video can disappear immediately, and is no longer accessible to the audience (Cai & Wohn, 2021), or, generally, even to the platform itself, depending on platform architecture, storage limitations, and other factors. This characteristic not only impedes the ability for prospective future investigations by law enforcement in egregious cases (Horsman, 2018), but also makes it more difficult for the trust and safety employees and community moderators of a platform to investigate and properly sanction the offending users as needed.

A related second affordance is *speed* — when violative activity is going on in-real time, the response from a platform needs to be extremely rapid to prevent the ongoing dissemination of the content in question. Even in a more benign livestreaming context (say a stream where a popular gamer is playing online chess with thousands of viewers), live
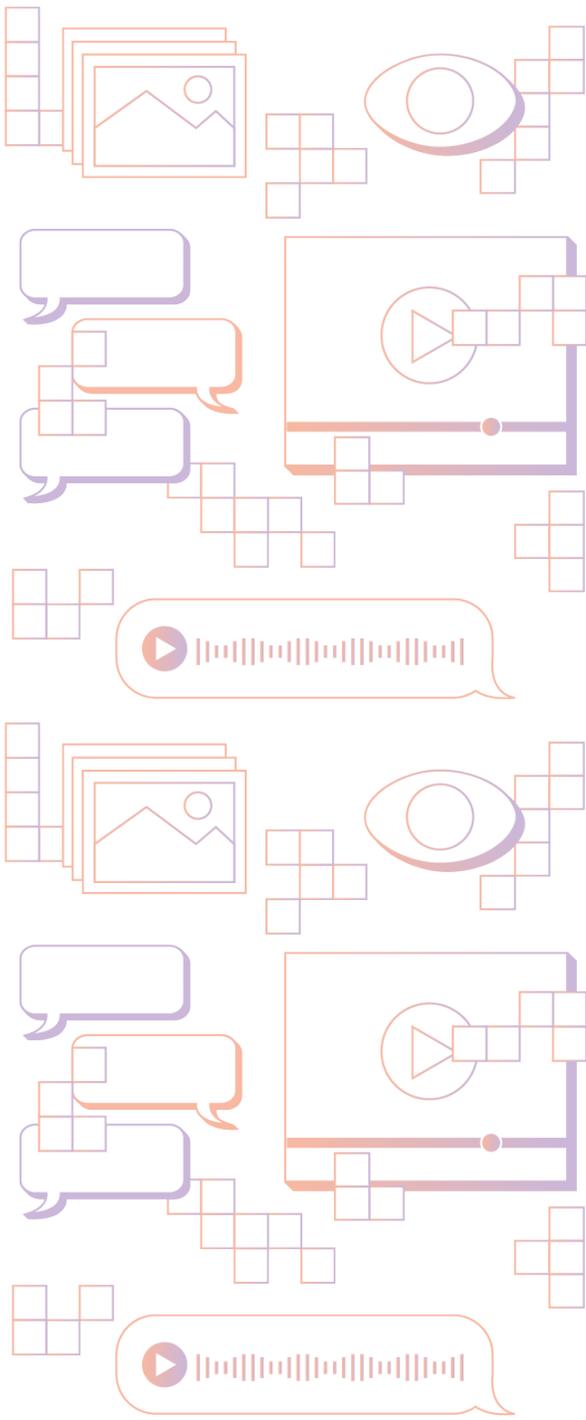
interaction and an active audience mean that *ex post* governance interventions are less effective and also potentially more difficult to implement (Cai & Wohn, 2019). An active body of research in human-computer interaction has documented the various tools and strategies that content producers use to moderate their streams, including under difficult conditions, like "brigades" of users seeking to harass certain streamers, as the streamer and affiliated volunteer moderators attempt to block those accounts and delete their comments (see Brewer et al., 2020; Cai et al., 2023; Xiao, 2024). In a context where there is a threat to public or personal safety, such as live acts of violence or self-harm, T&S responses need not only be quick, but also able to escalate into off-platform action involving law enforcement, public health officials, or other actors (Peralta, 2023; Zornetta & Pohland, 2022).

Finally, livestreamed media is, to at least a certain degree, *technically inscrutable* for platform trust and safety operations. Most livestreams by definition contain video and audio, and thus are inherently more difficult to analyze than classic text or image based posts. While industry has long sought to develop automated means for analyzing a range of multimedia, including video, for potential policy violations at scale (Cobbe, 2021; Gorwa et al., 2020), these techniques were developed to be deployed "at rest" on stored instances of content rather than on real time and constantly changing material (Shenkman et al., 2021). Analyzing content is inevitably far more complicated for end-to-end encrypted services such as peer-to-peer private video calling or conferencing (Kamara et al., 2021). Overall, this makes governing livestreaming a "routine resistant" problem for platform firms (Gorwa & Veale, forthcoming): one where the affordances of live surfaces arguably require them to develop specialized techniques that go beyond their usual operations on "cold" or "in situ" material.

While there is no existing published systematic overview of trust and safety practices in the livestreaming context, our survey of this landscape reveals a large number of related tools and practices that have been deployed or are in active development. Some of these are new, specialized strategies and systems that have been developed by some firms or vendors to specifically deal with CSEA. Others are merely new implementations of classic techniques generally used to detect sexual content or nudity, or to prevent spam, fraud, and other forms of abuse. Overall, these various interventions can be divided into three broad categories, which can occasionally overlap but nonetheless provide a helpful structure for thinking about platform interventions: "design" oriented modes of structuring certain service features against various potential forms of abuse, "content analysis" methods looking at the actual content of a stream, and "signals" methods that use metadata and behavioral account patterns to motivate moderation decisions rather than looking at the actual content. A summary is available in Table 1.

Overall, these various interventions can be divided into three broad categories, which can occasionally overlap but nonetheless provide a helpful structure for thinking about platform interventions: "design" oriented modes of structuring certain service features against various potential forms of abuse, "content analysis" methods looking at the actual content of a stream, and "signals" methods that use metadata and behavioral account patterns to motivate moderation decisions rather than looking at the actual content.

T&S Tools and Practices	Details
Design	Setting guardrails that require users to jump through additional hoops before they can livestream and thereby seek to reduce the likelihood of streamers disseminating illegal material
Account Thresholds	Requirements to meet certain criteria (e.g., account must be a certain number of days old, or must have a certain number of subscribers) to begin live streaming
Age Verification	Requiring users whose accounts go live to verify their age in some way; this can be very soft (e.g., self-declaration of age) or hard (mandatory upload of government ID)
Additional Verification	Requiring users whose accounts go live to verify some aspect of their identity by requiring more than just an email account, such as a phone number or credit card
Content Analysis	Automated content detection systems that analyze the content (e.g., video, audio, or text) of a livestream
Matching: Non Live Surfaces	A detection system that attempts to apply the industry standard CSA-hashing technology to non-live content (e.g., profile pics, backgrounds, other user uploaded content)
Matching: Sample Based Hashing	A system that takes sample frames from live streams and then runs them through established CSAM matching tools to try and detect "known" CSEA material
Predictive: General Safety Classifiers (Video)	Classifiers seeking to detect the likelihood that a stream (or a frame within a stream) contains content infringing a rule, not limited to CSEA (e.g., nudity classifiers, age estimation)
Predictive: Specific CSA Classifiers (Video)	Classifiers seeking to detect the likelihood that a stream (or a frame within a stream) contains CSEA content in particular, perhaps through a tuned version of a adjacent sexual content classifier, a combination of classifiers (e.g., sexual content detection and predictive age categorization), or a specially developed system
Predictive: Audio Classification	Seeking to detect abuse patterns in audio waveforms
Predictive: Text Analysis	Transcribing audio of livestreams and using it to run predictive risk classifiers
Chat Text Analysis	Predictive analysis of comments posted on a livestream; keywords and other potential indicators for flag
Signals	Forms of manual or automated analysis that use non-content data to track suspicious accounts or forms of behavior
Account Behaviour Indicators	Signals and metadata sharing around confirmed or predicted "risk actors"
Response Prioritization	Using signals or user flags to respond to potential policy violations in streams

▲   **Table 1:** Summary of common examples of industry tools and practices to address CSEA in livestreaming.

Design: Verification, Authentication, and Community Moderation Tools

Digital service providers of all forms have long implemented various general-purpose measures to counter fraud, spam, and other forms of potential abuse (Brunton, 2013). Techniques like two factor authentication (2FA, requiring prospective users to link their account to a working phone number, for instance) are widely used by social networking sites not only to make it more difficult to create fraudulent "sockpuppet" accounts, but also as a potential protection against phishing and account intrusion attacks (Tirfe & Anand, 2022). It is at least conceptually possible, if not empirically proven, that those seeking to create or disseminate various forms of underage sexual content may be less likely to do so if their account is linked to a phone number or other form of supplementary verification, although committed bad actors can evade 2FA requirements through disposable phone numbers, SIM swapping, and other techniques.

Similarly, various mundane-seeming design features — many of which are outsourced to streamers and their appointed moderators who can choose to implement various forms of chat filtering, "automoderation", or viewer friction — can play a role in structurally combatting multiple forms of abuse. For instance, Twitch streamers can set "channel level verification requirements", blocking viewers who have not verified their emails or phone numbers, or met other conditions (Twitch, n.d.). On certain platforms, streamers may be able to block link shares, and access blocklists of certain words or real time "chat toxicity" detection and blocking of various forms — some of which are community managed, while others are tools provided directly by platforms for easy integration and deployment.

Another type of threshold implemented by some firms on their live surfaces involves measures relating to an account's popularity. YouTube, for instance, had previously implemented a policy that all accounts seeking to go "live" needed to have at least 50 channel subscribers, a simple form of friction intended to prevent a bad actor from impulsively creating a YouTube channel and livestreaming an act of violence for example. Per our industry interviews, these general account thresholds were primarily rolled out to prevent livestreamed self-harm and other acts of violence (e.g., the Christchurch shooting), but also make the broadcasting of most forms of sexual content depicting minors substantially more difficult.

YouTube's threshold was widely reported in influencer marketing blogs (Tommy I., 2023), but it is not clear whether this applies only to mobile live streams (Google, 2024a). There are additional safeguards in place around livestreaming however: the latest YouTube documentation states that to go Live, one needs to authenticate a phone number, and for "advanced features" such as embedding livestreams, one must "build up sufficient channel history" or complete a personal document identification or biometric identification process (Google, 2024b). On TikTok, while there appears to be no similar threshold to go Live (TikTok, 2024e), there are thresholds for access to specific tools such as Live Studio (e.g., U.S. users must have at least 10,000 followers) (TikTok, 2024c).

SPECIFIC CHILD SAFETY MEASURES

Certain livestreaming platforms are increasingly also creating specific age-related thresholds intended to keep young people from violating their policies in various ways. TikTok community guidelines explicitly note that users under 18 cannot host streams using the platform's TikTok LIVE feature, while for Twitch, users must be 13 years or older (TikTok, 2024e; Twitch, 2024). Due to the US's Children's Online Privacy Protection Act of 1998, online services face special rules around how they handle the data of individuals under 13 years of age (Reyes et al., 2018). The law helped give rise to the dominant "self report your age" clickthrough form of internet account signup, but increasingly platforms appear to also be engaging in additional measures to predict the ages of some of their users. Twitch, for instance, uses undisclosed indicators to train models attempting to "catch and terminate accounts belonging to users under 13, as well as to block users previously suspended for being under 13 from creating new accounts" (Twitch, 2022). The company has also implemented some measures where accounts classified as "potentially vulnerable" through these kinds of predictive metrics need to go through "mandatory phone verification requirements' before being able to livestream."

TikTok similarly states that they use a set of undisclosed "additional approaches to identify and remove suspected underage account holders", looking at account behavioral patterns to make predictions about users under 18. Users flagged by these predictive models are required to undergo a mandatory age verification process if they wish to livestream, which involves sharing "a selfie [...with] government-issued ID and (b) a piece of paper clearly and legibly stating a unique code" (TikTok, 2024a). TikTok's transparency report states that the company removed on average 21 million accounts as suspected under-13s in the first two quarters of 2024 through automated techniques, although it does not disclose how many accounts were flagged for this kind of secondary age verification in the specific context of attempting to livestream (TikTok, 2024d). Our interviews with industry suggested that firms use the social graph of users, the types of topics and streams that they view, and other undisclosed forms of metadata to build these age-classification models, but there is little concrete public information about how comprehensive and effective these efforts actually are.

Identity verification systems also pose major privacy and security risks: there have been long standing concerns about the potential for third-party verification providers to become vulnerable points of failure and/or targets for hackers as they increasingly take hold of sensitive personal data.

The increasing move towards certain forms of age verification in the public livestreaming context has some commonalities with the legal requirements in the adult content industry. U.S. federal statute 18 U.S.C. 2257 requires US-based pornography sites to keep records confirming that their performers are all over 18. These types of requirements have been broadly implemented by many international adult content hosts as well: for example, the popular Cyprus-based camming site Stripchat, designated a Very Large Online Platform by the Digital Service's Act's special requirements by the European Commission in December 2023, has a rule as of summer 2024 that all new performers must submit passports, national ID, or other identification in order to show that they are over 18 before being allowed to create content (Stripchat, 2024).

In our interview with a large adult site, staff highlighted how they engage in third-party verification measures (using tools from a large third-party provider) of all performers, including on subsidiary sites that have livestreaming functionality. To help ensure that streamers depicted were in fact the ones that had gone through the government ID age verification process, this specific firm claimed to have human moderators "always in the loop" actively watching each stream for policy violations and checking performer identities against documents on file.

In the public livestreaming context, these forms of identity verification may make the proliferation of some — but not all — forms of CSAM more difficult. Accounts can be hacked and taken over by bad actors, documents can be forged, and the automated systems for age verification used by platforms and their third-party "age assurance" contractors make mistakes and raise significant privacy concerns. Identity verification systems also pose major privacy and security risks: there have been long standing concerns about the potential for third-party verification providers to become vulnerable points of failure and/or targets for hackers as they increasingly take hold of sensitive personal data (Blake, 2019; Persson, 2024). A 404 Media report in the summer of 2024 reported that one of TikTok's verification partners, an Israeli-based firm called AU10TIX, improperly secured its databases, allowing cybersecurity researchers to access the drivers licenses and other personal information of those required to use the firm's services (Cox, 2024b).

Content Analysis: Scrutinizing Streams for Policy Violations

Livestreams are inherently difficult to analyze for content policy violations with automated tools that work accurately and effectively at scale. That hasn't kept industry from trying various strategies, however. These range from classic approaches for finding previously confirmed "known" CSAM adapted to the real-time video context to more complex predictive methods seeking to accurately unearth "new" and previously unknown material with computer vision models. Industry also appears to be creatively implementing various matching and predictive tools during streams on supplementary material not directly part of the streamer's feed.

"KNOWN" MATERIAL: VIDEO AND AUDIO

Little has been publicly disclosed about the specific techniques that platforms are using to work with live video when it comes to the detection of potential CSAM content. Firms like TikTok mention that they may use a wide basket of technologies, "including [their] own systems and hash-matching software like Microsoft's PhotoDNA, Google's Content Safety API, and YouTube's CSAI Match", without going into depth on the specific implementations and relative advantages of these various systems (TikTok, 2024b). In interviews, some industry participants mentioned that techniques like "scene sensitive video hashing" (SSVH) can be used to collapse videos into their "key frames", which then can be run through conventional hash-matching systems like Photo DNA.

A small number of firms appear to also be deploying even more computationally lightweight techniques, such as audio hashing, which have become relatively robust in recent years and have gained popularity through "what is this song" matching tools like Shazam. Here, industry can make use of audio hashes of confirmed CSAM, to, as with SSHV approaches, prevent streamers taking advantage of live offerings to publicly "loop" previously confirmed CSAM content. Although these technologies are generally perceived to be fast and accurate, the worry of some interviewees was that it was not worth the computational cost for many firms given the perceived rarity of this kind of "previously known" CSAM being broadcast publicly by streamers.

Where firms do appear to be actively using hashing technologies, according to our interviews, are instances in which content is static, at-rest, and can be scanned and matched for potential violations. For instance, large scale streaming platforms that allow users to upload backgrounds, images to the chat, custom emojis, or other types of media consistently hash those uploads to prevent the easy dissemination of known CSAM. Hashing unencrypted and non-live content is widely deployed by leading platform services with varying business models (peer-to-peer messaging, cloud, user generated content upload) as a general safeguard and best practice.

"NEW" MATERIAL: VIDEO, AUDIO, TEXT

Some third-party vendors are offering products which seek to allow platforms of various sizes to integrate these kinds of video analysis solutions into their livestreaming environments. One tool being marketed by a large Trust and Safety solutions firm, for example, pulls consecutive frames from a live video and then runs image classification on them. Various predictive models then return scores for assorted potential 'harm areas', allowing clients to take moderation actions based on specific scores or combination of scores.

More active efforts are also underway to analyze content in the unique livestreaming context using new techniques that could potentially help platforms identify "previously unknown" material — content being created live on the spot that by definition has not been seen before and thus would not be contained in hash databases of known CSAM. The natural starting point for this kind of analysis is to try and work with the video stream itself, using various computer vision models to predict the possibility that certain policy violations are ongoing. Such classifiers are increasingly being used by large firms in adjacent areas such as general sexual content detection: for example, in December 2023 an Instagram press release noted that the company had deployed a "new automated enforcement effort" that increased their "automated deletions of Instagram Lives that contained adult nudity and sexual activity" fivefold (Meta, 2023).

Some third-party vendors are offering products which seek to allow platforms of various sizes to integrate these kinds of video analysis solutions into their livestreaming environments. One tool being marketed by a large Trust and Safety solutions firm, for example, pulls consecutive frames from a live video and then runs image classification on them. Various predictive models then return scores for assorted potential 'harm areas', allowing clients to take moderation actions based on specific scores or combination of scores. A popular product with public documentation and similar features is Amazon's Rekognition, which offers nudity detection in images and video and also allows clients to fine-tune it for custom purposes (Amazon Web Services, 2020). To be able to potentially detect sexual content featuring minors, one strategy discussed by vendors involves combining age-estimation tools with various publicly available nudity or sexual conduct classifiers (of adults). The ensuing models are trained using supervised or unsupervised learning methods to try to detect new content based upon patterns from that training data.

Another strategy being deployed by a newer crop of child safety-focused vendors involves creating classifiers specifically for detecting underage sexual content, training models on datasets of confirmed child abuse material that are provided by law enforcement or organizations like NCMEC or the Interent Watch Foundation. The idea here is to use datasets of CSAM to predict the likelihood that a new image (or video frame) may also be CSAM. Publicly advertised tools with this kind of functionality include Thorn's 'Safer Predict' (Thorn, 2022) and SafetoNet's 'HarmBlock' product (MSAB, 2023), the latter of which claims to be able to work in a livestream context (Payt, 2024). Advocates argue that this kind of approach is more promising than bundling problematic age-estimation models with nudity models, although there are open questions about how globally representative the datasets of CSAM these models are, the ethical issues around making these datasets available to private companies, and their actual efficacy at platform scale given the lack of public benchmarking and independent efforts to verify these models in various settings.

Another emerging technique involves moving away from the actual video material and instead working with the audio using predictive models. Some interviewees mentioned their efforts to train predictive models on audio, arguing that videos of child sexual abuse exhibit distinct waveforms and patterns that can be used to then find similar videos. However, it is not clear how robust and distinct these waveforms and patterns are, especially when seeking to detect infringing behavior on platforms that also host ample depictions of violence and sexuality in, for example, popular video games.

The more common approach appears to work with the transcribed audio of streams and text associated with streams (e.g., in the chat and other stream metadata). Many large platforms have already bundled speech-to-text captioning models into their live offerings (allowing for multilingual live closed captions, for instance), making it relatively simple for them to use these transcripts for other types of moderation actions. Text analysis techniques of varying complexity (ranging from simple keyword-based "flag lists" to tuned large language models) were discussed with us by the employees of public-facing livestreaming platforms as a tool increasingly used to flag suspicious streams for human review. Not everyone is doing this, however: some participants mentioned that they wished that their firm would invest in transcription at scale, as it would unlock a lot of new trust and safety possibilities for them, but that doing so was seen as prohibitively expensive from a compute and storage perspective.

There are of course risks in using machine learning models for textual analysis including missing the context in which a post is made, particularly in multi-cultural settings; and exacerbating discrimination against already marginalized groups (Duarte et al., 2017). More recent advances in multi-lingual large language models that can be used to assess content in different languages still pose problems for analysis in "low-resource" languages, or those for which training data is scarce (Nicholas & Bhatia, 2023).

Working with transcripts or audio waveform signals can complement the classic "flagging" model of community guidelines violation reporting by ordinary users, long a staple of how intermediaries seek to identify and respond to complaints (Crawford & Gillespie, 2016). Virtually all platforms publicly state that they seek to respond to flags in livestreams as quickly as possible, although they increasingly have begun to use additional predictive tools and forms of metadata analysis to help prioritize this review so that they can respond to the most 'urgent' and high-stakes reports.

One way this can be done is by creating a "risk score" of the kinds of interactions in a chat, and/or by analyzing other forms of information provided by a streamer that aren't necessarily part of the core stream (for example, video thumbnails, titles, and other keywords), or by using audio or video-based analyses as flags for human observation. Thorn advertises a predictive text classifier that does this kind of risk-scoring, and their public marketing materials feature a quote from a service provider claiming that the company's "text classifier significantly improves our ability to prioritize and escalate high-risk content and accounts. The multiple labels and risk scores help our team focus on problem accounts, some of which we had been suspicious about but lacked actionable evidence before we deployed the classifier" (Thorn, n.d.). This kind of text classification can in theory be used by firms on transcribed live streams to flag them for human review.

Text analysis approaches are also being used by certain firms to try and predict the incidence of "grooming" and CSAM-affiliated conduct, rather than solely actual CSAM: for instance, Thorn's Safer product claims that it can search for "child sexual exploitation behavior" in text (Thorn, 2022). While firms are publicly vague about specifics, it seems as if multiple platforms are developing lists of keywords that can be used as indicators of child safety issues broadly construed, despite the high potential degree of false positives and the potential issues across different linguistic contexts. The Tech Coalition, together with Thorn, has set up a "CSAM Keyword Hub" allowing for platforms to access and communally manage a set of keywords in multiple languages that can be used for various trust and safety interventions (a short public FAQ states that the hub should ideally not be used "for strictly blocking specific keywords that match the list", as "the strong preference is to use the list to kickstart the training of machine learning models" that can flag material for review or block items from a chat) (Thorn, 2024).

As one industry interviewee put it, keyword oriented systems can be brittle and yet nonetheless helpful: even though their safety team is frequently inundated with flags that pertain to streams that are clearly not problematic when, for instance, active fan discussion of a DC Comics film "Suicide Squad" leads to the sudden proliferation of streams with self-harm keywords in the title and description, text-analysis systems can still be used to analyze public chats, stream descriptions, and stream titles to guide ensuing human review.

Signals: Investigating, Tracking, and Predicting Violative Behavior

One industry expert noted that efforts to prevent CSAM in livestreaming have become more sophisticated in recent years, and have shifted focus away from detecting the potentially infringing content and towards understanding the behavior of confirmed or suspected bad actors. From this perspective, a first wave of trust and safety efforts sought to parse stream video, a second wave has moved towards using audio (and especially transcription), and the latest cutting-edge of industry practice involves a third wave of interventions primarily based on "signals" of actor behavior. The latter set of practices is the result of a multi-year effort to try and develop non-content oriented measures for CSAM detection (e.g., by looking at file metadata) (Pereira et al., 2023), and while not initially developed for the streaming context, these tools can also be a helpful part of streaming platforms' trust and safety operations.

One way in which signals can be used is to share data about accounts and activities that violate a platform's policies regarding CSEA. This can undermine the ability of known bad actors to operate across multiple platforms. A second approach is to utilize signals to predict whether certain activities or accounts are potentially involved in the distribution of CSEA. Most major livestreaming platforms will employ both approaches.

A major new project facilitated by the Tech Coalition consortium that supports signal based approaches to preventing the distribution of CSEA is Lantern. Participating platforms can share "signals", such as metadata linked to accounts that have been confirmed to be active distributors of CSAM, for use by other platforms. These indicators — examples provided by Lantern include email addresses, usernames, and certain keywords — can be used to inform investigations made by specialized teams at the platforms, or to feed models which can be used to flag (or remove) accounts proactively (Tech Coalition, 2023). Meta's ThreatExchange platform also allows participants to share and access signals in a structured way, and their public documentation includes numerous potential signals, some of which can be very granular, such as information about a user's browser fingerprint ("user agent string"), IP address and associated latitude and longitude, or the name and address revealed in a "Who Is" lookup for a web domain associated with an account (Meta, 2024b).

As one participant put it, the idea is for firms to engage more actively in reducing the ability of their platform to be used for CSAM dissemination, not only engaging in a detect and report mode but also, aspirationally, towards a predict and disrupt model of trust and safety.

Livestreaming platforms can utilize signals whether or not they are not part of such industry sharing efforts. For example, if the patterns of behavior around a certain stream are highly unusual (e.g., as one expert put it, the stream was just instigated by a brand new account and suddenly has a large audience all coming from an external link, with mostly new accounts in the chat, most of which are linked to IP addresses associated with VPN servers and have emails linked to anonymous or privacy-preserving email providers), these factors can help categorize a stream as high-risk for rapid moderator review. Some firms stated that signal approaches are already helping them make more informed decisions around the moderation of certain accounts and are generally helpful from a trust and safety perspective before a human review can occur. Signals open up a whole range of tools in the moderation toolbox, which can go beyond simply removing content: if certain user activity is scored as potentially risky, firms can implement friction on ongoing streams, including lowering the bandwidth on the stream, increasing latency on the chat, or even logging out stream audience members so that they need to log back in while a human moderator conducts an investigation.

This is an emerging industry practice which has yet to be subject to extensive academic, civil society, or journalistic scrutiny. Firms and experts both suggest that this is a highly promising approach in terms of efficacy, allowing firms not only to seek to robustly prevent bad actor accounts from sharing content (or removed accounts from returning to the platform), but also potentially to use during streams in combination with other techniques to make more accurate moderation interventions.

As one industry participant put it, the most proactive firms have increasingly shifted away from seeing their efforts to safeguard livestreaming as a content moderation issue, and instead seem to think of it as a 'threat intelligence' problem with ties to the broader field of cybersecurity and fraud prevention. This turn is evident in our interviews not only with public-facing livestreaming platforms, but video calling services as well, where firms can decide to make interventions based not on private communications, but rather onsurrounding metadata. As one participant put it, the idea is for firms to engage more actively in reducing the ability of their platform to be used for CSAM dissemination, not only engaging in a *detect and report* mode but also, aspirationally, towards a *predict and disrupt* model of trust and safety. Previous research also noted that an emphasis on metadata as opposed to analysis of actual user generated content was important for platforms that provide end-to-end encrypted communication services (Kamara et al., 2021).

Some industry participants mentioned that they believed that signals were a powerful tool to help firms be more selective in the use of more potentially harmful or privacy-invasive techniques — such as the forms of content analysis mentioned in the previous section. Rather than transcribing all streams and running a text analysis model on the outputs, in theory, a signals approach could allow firms to decide which streams should be scrutinized in further depth, reducing the chance of random false positives and minimizing the extent to which completely random users end up caught in these kinds of automated dragnets. Also, analysis of signals could address some of the shortcomings of content analysis using large language models to examine chats or audio transcripts in "low-resource" languages.

These approaches are still in their early stages. Nevertheless, our interviews also raised concerns that the bulk use of signals has evolved not just to suggest careful manual investigation by experts (e.g., platform's child safety investigations teams), as previously was industry practice, but increasingly to develop models that make automated decisions at scale and could lead to unforeseen consequences. It's not clear to what extent signals and metadata are being used to develop classifiers by companies.

In particular, more research needs to examine the impact of this highly opaque developing set of practices on populations who may happen to have some "suspicious" characteristics. Are privacy-conscious users (who are using VPNs or email providers other than Gmail or Outlook) being thus profiled and having their user experience reduced or their accounts mistakenly removed? What about the very broad demographic of users associated with certain global majority countries that could also theoretically be identified as suspicious (due to their IP addresses and location characteristics) and thus subjected to other types of investigation or content analysis?

III. Policy Implications of Existing Trust and Safety Practices

These three broad modalities of interventions against various forms of publicly broadcast sexual content involving minors all have different implications, failure modes, and human rights impacts for platform users in general and survivors specifically.

Limitations of Existing Design Interventions

Structural and design based interventions that set popularity thresholds before an account can livestream likely prevent many types of the most egregious abuse without needing to deal with the privacy and security shortcomings of age verification tools. Although these are not perfect, popular influencers who either have had their account taken over or are breaking community guidelines with self-generated sexual content are more likely to be reported by viewers and removed. However, stream-first platforms which do not have other ways for accounts to build up an organic following (like they can, for instance, on TikTok or Instagram through their non-live content) do not have this luxury and thus need to look to other approaches.

An approach that requires only content creators to verify their age (rather than just ordinary users/stream viewers) may seem a proportional safety measure to some — after all, not being able to stream is not the same kind of barrier as not being able to participate at all in the digital public sphere — but effectively and securely implementing age verification measures is not only very difficult to do safely in general, but also particularly high risk when it involves the biometric data of young people (CNIL, 2022; Forland et al., 2024). These technologies can also create disproportionately more risks for specific groups of children such as those with disabilities (Bhatia & Aboulafia, 2024). Additionally, as one participant noted, approaches that, for example, require a credit card on file to curtail risk of abuse could render these forms of protection or signals a luxury good. In fact, because of the burdens they place on all users of a given platform, including disproportionate impacts of marginalized communities, legal mandates requiring the use of age verification technologies can restrict or chill legally protected free speech (Ruane et al., 2024).
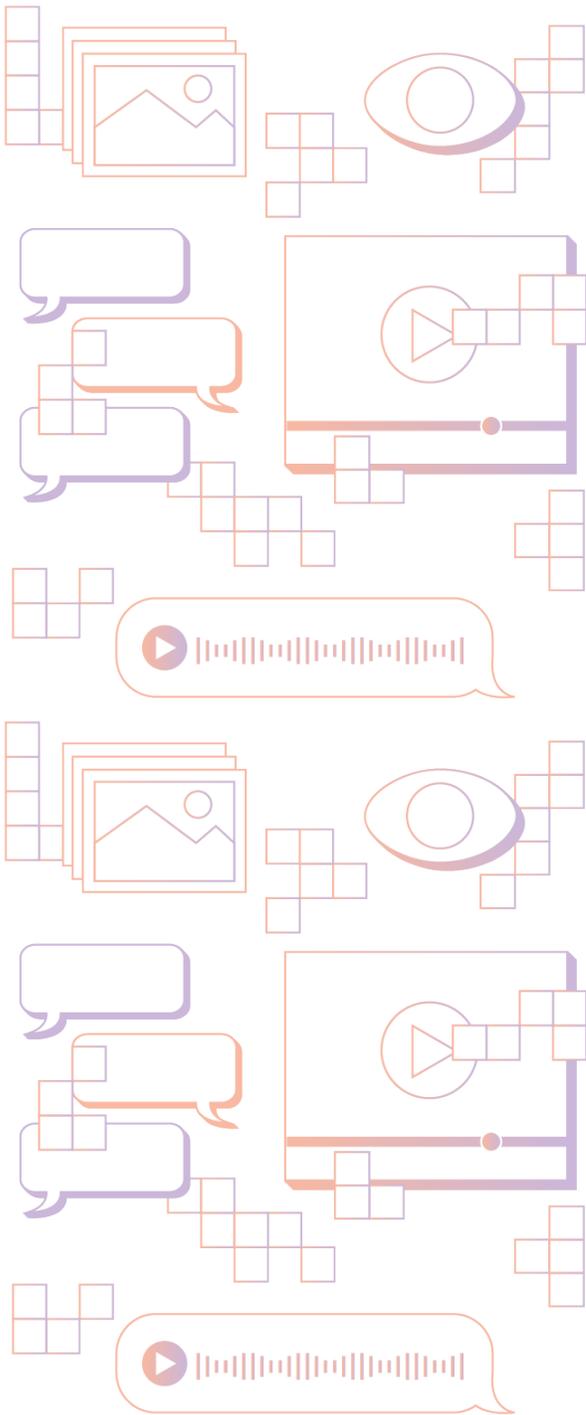
Ramifications of Predictive Livestream Video Analysis: Bias, Overblocking, Data Quality, and Data Sources

The international regime for reporting detected child sexual abuse material to NCMEC's CyberTipline, which is legally required under US law, and the ensuing "fingerprinting" of this confirmed content to be shared with firms for their detection efforts may be imperfect and could be improved (Grossman et al., 2024). Despite its issues it is nevertheless a relatively proven and established system for identifying known CSAM. The challenge for CSAM prevention in the livestreaming context is that the majority of potentially infringing content cannot be detected through this time-tested technique.

How is the training data actually sourced, and can consent be obtained before models (which may be proprietary and then sold as a service for profit) are trained with it? How can other stakeholders verify the performance of platform systems, including for key automated content analysis tools? Overall, the "new content" CSAM-prediction space faces a major underlying problem of benchmarking.

As industry continues to move towards other techniques to detect previously unknown CSAM — or is required to do so by legislation — the potential for false positives, and therefore for problematic and unintended downstream effects on ordinary users, increases significantly. Computer vision systems seeking to classify images (in this case, frames of live videos) are notoriously prone to misclassification, especially when presented with unexpected variation on racial and gender lines (Wang et al., 2022; Zhao et al., 2021). Supervised learning systems struggle out of domain, and can fail spectacularly in the real world when facing less-than-ideal conditions (such as low resolution video, poor lighting, obstructed figures or side profiles of individuals). Some of the core technologies underpinning industry predictive approaches, such as age estimation models, can exhibit troubling racial and gender biases, as well as general performance issues that critical scholars have used to call the use of these systems at all into question (Stardust et al., 2024). At platform scale, even purportedly accurate systems could yield many thousands of daily false positives.

Past CDT research has explored the prospective issues facing implementations of automated multimedia content analysis in depth (Shenkman et al., 2021). Although the focus of that analysis was domain-general, rather than targeted to the CSAM prevention context, the same concerns relating to these methods — and issues of robustness, data quality, lack of context, poor measurability, and explainability — all apply here as well. Most of the engineers and technically oriented product managers in industry we spoke to were candid about the limitations of their technology, noting that these tools in general are imperfect, and that certain levels of bias are inherent to predictive machine learning models seeking to do image classification. Vendors, however, are more bullish, sensing a commercial opportunity and seeking to refine their technology on the fly once it has already been deployed.

There are additional related issues to data quality, data provision, and ethics that are unique to the CSAM detection and prevention context. How is the training data actually sourced, and can consent be obtained before models (which may be proprietary and then sold as a service for profit) are trained with it? How can other stakeholders verify the performance of platform systems, including for key automated content analysis tools? Overall, the "new content" CSAM-prediction space faces a major underlying problem of benchmarking. There are currently no public performance metrics through which firms could test the accuracy of their systems or through which experts, policymakers, and researchers could better gain an understanding of their efficacy, as well as the extent of what is really possible (Laranjeira da Silva et al., 2022).

A key issue is how to train models that are intended to detect previously unknown CSAM, particularly given that CSAM is unlawful and therefore not readily available to be used as training data. A growing number of vendors have in recent years entered into contracts to develop investigative tools for law enforcement. An example is the multi-organization Project ARICA, funded by the European Commission (ARICA, 2023). With complex contracts, legal carveouts, and data sharing arrangements, our interviews touched on models specifically trained for image classification with government-held datasets of confirmed CSAM. Most of these are to aid law enforcement investigations, but some of these models appear to be marketed for commercial use. There generally is no public documentation associated with these products, but our conversations with vendors involved recurring claims that their tools can accurately predict the age of youth and identify sexual image frames within streams, providing accurate estimations without major racial or gender bias.

Without any openness, information about testing, or standardized benchmarking, it is difficult to evaluate the efficacy of these claims. There are many reasons to be skeptical, however. As one engineer put it, image classifiers looking for "new" CSAM content are fighting an uphill battle: to work well at platform scale, the systems need to be "99.99999999% effective", or else they will be mistakenly flagging thousands or even tens of thousands of pieces of content per day. If not integrated into the right kind of enforcement pipeline, this will lead to erroneous takedowns and unforeseen patterns of content suppression, potentially with the largest effects on legitimate forms of sexually-tinged expression. On one hand, this is a particularly high-stakes area with major implications for victims of sexual abuse, and some interviewees expressed concern that being too careful about human rights impacts would disincentivize firms from developing new and innovative technologies that could help rein in certain forms of physical violence. Nonetheless, there are still worries that firms are not or will not robustly oversee these models and implement them with the appropriate level of care: for instance, a public quote displayed as an endorsement by one major safety vendor brags that their client "[doesn't] even bother reviewing the content [the tool] flags — it's that great and consistent" (ActiveFence, 2024).

There are many reasons to be skeptical, however. As one engineer put it, image classifiers looking for "new" CSAM content are fighting an uphill battle: to work well at platform scale, the systems need to be "99.99999999% effective", or else they will be mistakenly flagging thousands or even tens of thousands of pieces of content per day

One major challenge relates to data quality, and the ability of models to achieve accurate domain expertise in the specific problem they are intended to solve. The classic computer science problem solving approach to identifying content X in the wild involves collecting large amounts of content X and then deploying various methods to tune or train an image classification model on patterns in that data. Theoretically, the ideal "new CSAM" detection model would be trained on some kind of globally collected dataset with contributions from law enforcement and partner organizations to achieve a globally representative set of training data; these models would be then tested on a robust multi-jurisdictional benchmark dataset filled with not just completely unrelated images (e.g., from ImageNet or other generic image classification datasets) but also with many legitimate images depicting young people and sexual but not illegal content. Barring these kinds of collaborations, or other related measures — audits of datasets by independent experts with the requisite skills and training, model benchmarking best practices and/or red-teaming exercises done by third-parties, and other forms of testing and transparency gated for key stakeholders — it is extremely difficult for researchers or policymakers to ascertain whether these products are more than mere AI snake oil.

A few participants mentioned the challenge of ethical data sourcing. There is no current standard best practice for obtaining consent for the further use of victim data (e.g., to train models; see also Laranjeira et al., 2022). Civil society advocates we interviewed expressed reservations about actors using CSAM to train models and then selling access to the resultant products at a profit without the consent of those depicted. Industry conversations also highlighted the challenges of navigating an environment where the data at hand is extremely sensitive and legally contingent, creating difficulties in collecting and handling it. Nonetheless, some vendors suggested to us that they may not only getting legal access to sensitive CSAM material from law enforcement or their partners, but are also actively searching it out in illicit fora and 'the dirty parts of the internet," pitching that as their core value added, although these claims are difficult to independently verify.

Issues with Signal Sharing: Privacy, Opacity, and Little Recourse

Experts in industry and academia are hopeful about the ongoing move towards use of signals, and away from actual analysis of livestreams across a range of different platform types. According to proponents, this development — which is still in its early stages — can not only lessen the possibility of false flags and false takedowns made by imperfect automated media analysis systems, but also should be a far more effective way to block and investigate the relatively small numbers of motivated bad actors.

Beyond just the livestreaming context, the general use of such open-ended measures across different platform surfaces (e.g., non-encrypted direct messages) poses a major potential freedom of expression issue if it leads to automated interventions against users or even identifies certain groups for in-depth investigation.

The challenge, however, is that the way signals are used is opaque and may lead to biases and failures that are hard to detect internally (inside firms) and externally (by potentially affected members of the public). Some potential signals are far less concrete and less auditable than MD5, PDQ, or other hash signatures. For example, the Lantern project's list includes "keywords used to groom", an inherently fuzzy concept (Tech Coalition, 2023). In addition, there are questions about the extent to which different firms can properly understand the techniques of data production and model training used to develop these classifiers. The opacity around the use of signals, particularly in how they are used to predict bad actors or behavior whether in automated systems or not, means that it is difficult to independently verify their efficacy outside of industry statements.

Use of signals could also have significant privacy implications. Civil liberties oriented activists in the EU are already mobilizing against the use of contact-oriented text analysis tools by platforms. Pirate Party Member of the European Parliament Patrick Breyer filed a lawsuit in early 2024 with a local German court on the suspicion that such "grooming" related text-classifiers were being deployed on his Facebook Messenger chats (Breyer, 2024). The European Commission's draft regulation on the prevention of child sexual abuse has featured extensive discussion of predictive "anti-grooming" technologies and likewise been mired in controversy and wide-spread organization against the proposal from civil society (EDRi, 2023). If signals-oriented approaches are being used on non-encrypted surfaces which nonetheless come with some expectation of privacy — such as private messages, as alleged by Breyer — there is an additional worry that people's innocuous everyday chats will lead to them being mistakenly caught up in these kinds of automated dragnets.

Beyond just the livestreaming context, the general use of such open-ended measures across different platform surfaces (e.g., non-encrypted direct messages) poses a major potential freedom of expression issue if it leads to automated interventions against users or even identifies certain groups for in-depth investigation. In the US, for example, the REPORT Act now requires platforms to report instances of "enticement", leading to what Riana Pfefferkorn has described as "increased incentives to overreport," and the possibility that "innocent online speech — say, flirting between two teens, or a user quoting song lyrics about pimping — gets reported, first to NCMEC and from there to the police" (Pfefferkorn, 2023).

With initiatives that attempt to enable signal sharing across industry these problems could become exacerbated along with the introduction of new concerns. For example, mistakes, biases, and other data-related issues (or decisions made about parameter fine-tuning, etc) that lead to problematic outcomes could, perhaps without it being realized by firms in the same consortium, proliferate across platforms, leading to wider effects across the platform ecosystem and embedding issues relating to explainability and system opacity even deeper into industry trust and safety workflows.

Consortia like Lantern pose some similar challenges to those of systems like the Global Internet Forum to Counter Terrorism's (GIFTC) hash-database. For instance, if someone's account is hacked and then used to disseminate CSAM, and then their account indicators (username, IP address, other associated metadata) are shared with other firms, it is possible that the effects of stolen or leaked credentials on one platform have a major and hard-to-remedy ripple effect that leads to them being blocked across many other services as well. As with longstanding concerns with hashes and GIFCT, key accountability questions involve the ways in which material is incorporated into the consortium, the way in which it is overseen and/or audited by experts, and the broader recourse structures in place (Llansó, 2020).

Privacy concerns can be exacerbated to the extent personal data is contributed by and shared across companies with these kinds of frameworks. For example, a combination of different meta-data can be used to infer personal or sensitive information including identities (Kamara et al., 2021). Also, while companies will put in place measures to prevent erroneous classifications they are still possible, and it is often difficult to be aware of these until someone has to seek recourse.

Some firms spoke to us about the ways in which they use "appeals" processes to allow users erroneously removed to try and get their accounts reinstated, and large swings in their account takedown appeal metrics indicate to them that their signals-based models have become overly sensitive, and perhaps are catching too much legitimate activity in their efforts to remove spammy, fraudulent, and/or potentially dangerous CSAM-proliferating activity. Well-structured appeals processes are thus crucial to provide users recourse and an ability to contest mistaken moderation decisions, but should not be the only safeguard that firms implement when relying on these kinds of signals-oriented trust and safety techniques.

Conclusion
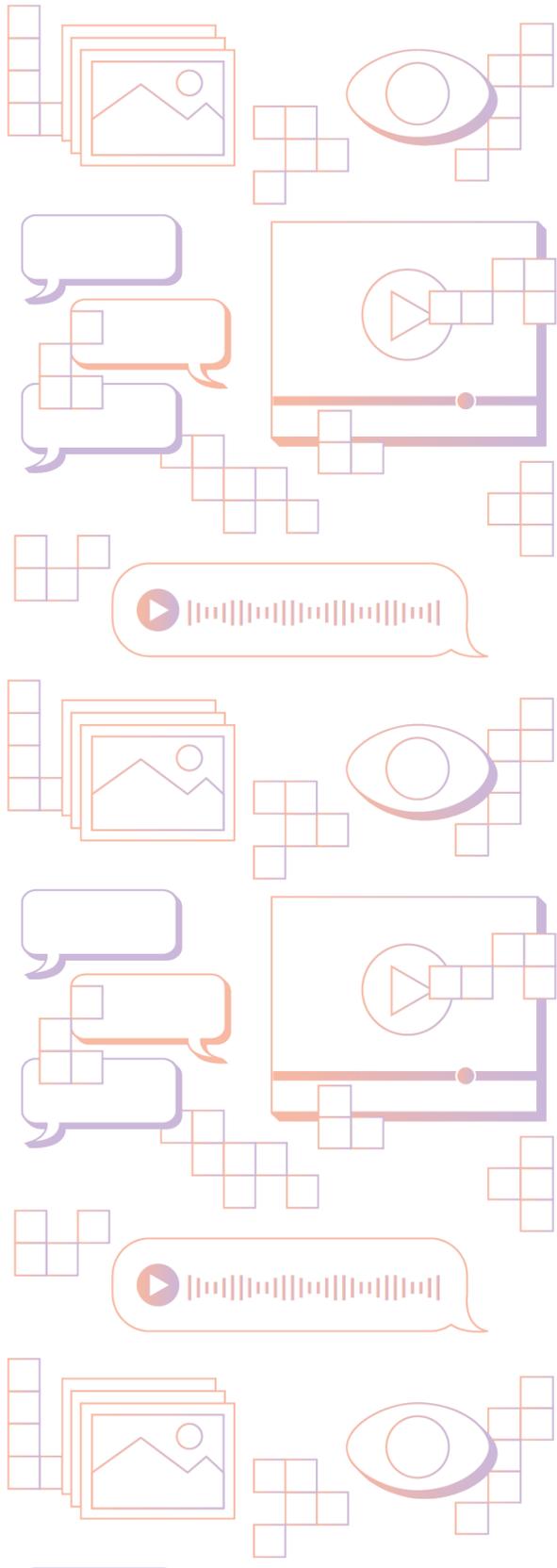
The specific strategies and systems for trust and safety that the platform industry is currently developing and deploying to more effectively prevent CSEA material from being disseminated via real-time video surfaces are continually and rapidly evolving. This is a complex landscape with little public documentation of best practices. It is also one at a critical juncture, as increasing international interest in policy interventions to address child safety and online harms has also led to a proliferation of "safety tech" vendors and other third-party groups pushing a wide range of technical solutions and bespoke "counter-CSEA" automated content analysis technologies.

Online service providers that operate real-time video surfaces of all kinds face heavy pressure to intensify their efforts to limit the proliferation of CSEA and related harms. The stakes are particularly high given the current transnational movement of child safety oriented platform regulation and "online harms" bills, efforts to more stringently control young people's access to certain information and online services (Marwick et al., 2024; Witting, 2019), and renewed efforts at "techno-legal solutionism" applied via purportedly "kid friendly" design and policy changes (Angel & Boyd, 2024). Policy fights centering around certain livestream adjacent platforms — especially porn sites and video calling services — are increasingly implicating broader topics like end-to-end encryption and user/age verification, themselves part of extremely consequential debates around cybersecurity, privacy, and digital autonomy for not only young people, but adults as well (Child Rights International Network & defenddigitalme, 2023; Forland et al., 2024; McKee & Lumby, 2022).

A path to take this conversation forward will involve continued engagement between firms, vendors, civil society, and academic experts, ideally with more open experimentation and engagement with the actual technical systems being proposed by these actors. There are several areas in which progress can be made in this regard:

Greater transparency is needed to help improve efforts to address CSEA on livestreaming platforms.

We urgently need mechanisms for better benchmarking practices for predictive image classification models, as well as potentially for high-stakes "signals" oriented "user risk" profiling models. As it stands, however, vendors have little incentives to submit their products to such open forms

of review, given that poor performance on industry-standard benchmarks would have a major impact on their ability to court clients. Some of the standard best practices from the leading edge of machine learning research have yet to trickle into this increasingly politicized and sensitive domain, but it is clear that at the very least some simple forms of dataset and model transparency — e.g., model cards and datasheets providing information about how models were trained (Crisan et al., 2022; Mitchell et al., 2019) could be incorporated by firms and vendors. If these are not done in a fully public manner, then in a gated manner where they could be accessed by key policy, civil society, and industry stakeholders. Sharing research sponsored or carried out by vendors could also be helpful. In several interviews, we were made aware of non-public white papers and reports that are available to prospective clients but not the public. While voluntary transparency frameworks do exist, see for example (Tech Coalition, 2022), the challenges we faced in conducting this research and trying to learn how industry trust and safety systems operate suggest that these frameworks are not enough.

The stakes are particularly high given the current transnational movement of child safety oriented platform regulation and "online harms" bills, efforts to more stringently control young people's access to certain information and online services

Vendors and livestreaming platforms should be explicit about the limitations of automated approaches to detecting and addressing CSEA and build their trust and safety systems accordingly.

Being cognizant of the inherent technical limitations of predictive machine learning models, platforms should continue to exert care about the ways in which these systems are deployed, building in adequate safeguards to try and mitigate bias and the suppression of legitimate expression. Using predictive models to prioritize streams for rapid review seems to be a reasonable measure, helping platforms identify when to involve a human reviewer and, hopefully, allowing them to make nuanced decisions based upon context. These reviewers should be fairly-compensated expert moderators with specialized training and psychological support designed to help them deal with potential exposure to CSAM and CSAM-related material.

Focus on design interventions that empower users including minors.

While many of the design interventions of major livestreaming platforms focus on identity verification and authentication, some efforts include tools for streamer and community moderation. The needs of streamers, including minors, to protect themselves from being targeted with or used to distribute CSEA are worthy of greater attention when it comes to design based solutions. For example, one design based approach that was not raised in our discussions with industry is to provide users, particularly minors, with the right set of tools and reporting mechanisms to help them protect themselves.

Reporting can be an important tool for children to address CSEA problems such as online grooming (Kennedy et al., 2024). Prior research suggests that, in the case of direct messaging platforms, having the ability to track the platform's response to a user's report is particularly important (Luria, 2023). This is also relevant for livestreaming platforms, and tools such as these could be customizable to account for the unique risk profiles of specific groups of minors (Luria, 2023).

Multistakeholder governance models can improve accountability of approaches to address CSEA on livestreaming.

Best practice frameworks around the implementation of these systems could be developed not only through the continuing work of organizations like the Tech Coalition, but also through critical multistakeholder engagement, in fora that involve child safety organizations actively engaged on a broader set of digital rights and civil liberties. Siloing conversations across different stakeholder groups in government and civil society is not a sustainable long-term model for these crucial policy discussions. Indeed, multistakeholder models applied to governance mechanisms around terrorist content, for example, have benefited from this approach and can be improved with more accountability (Bhatia, 2024). For example, these can allow relevant actors to inform the design of audits and other forms of external evaluations. This may be particularly relevant for signals-related measures. The Tech Coalition's initial launch of the Lantern project featured the news that they had commissioned an external human rights impact assessment of the initiative (Tech Coalition, 2023), and involvement from specialized third-party auditing and reporting firms should be welcomed going forward given the high stakes of the issue area at play.

Overall, we are at a key juncture for the future of addressing CSEA specifically on livestreaming platforms. Addressing this problem is critically important given the impacts on children, parents, and their communities, and so this is a hugely consequential and high-stakes area of platform governance. Vendors and industry alike are understandably eager to show that they are developing innovative new tools to handle stakeholder demands, and taking the general area of child safety and child sexual abuse seriously, but poor implementation (or poor trust and safety design, with systems that are fundamentally flawed) will decrease, rather than increase, policymaker and public confidence in platforms' trust and safety over the longer term. All stakeholders involved should therefore have an interest in ensuring that emerging industry practices are implemented in a careful, responsible manner that is informed by a realistic assessment of the prospective tradeoffs, technological limitations, and knock-on effects that these different interventions could have.

References

ActiveFence. (2024). *Real-Time Video Content Moderation*. ActiveFence. https://www.activefence.com/video-content-moderation/ [perma.cc/LP6B-N2LH]

Allyn, B., Goodman, S., & Dara Kerr. (2024, October 13). Inside the TikTok documents: Stripping teens and boosting "attractive" people. NPR. https://www.npr.org/2024/10/12/g-s1-28040/teens-tiktok-addiction-lawsuit-investigation-documents [perma.cc/JBN9-4DVL]

Amazon Web Services. (2020, October 12). *Amazon Rekognition adds support for six new content moderation categories | AWS Machine Learning Blog*. https://aws.amazon.com/blogs/machine-learning/amazon-rekognition-adds-support-for-six-new-content-moderation-categories/ [perma.cc/A72N-T26W]

Angel, M. P., & Boyd, D. (2024). Techno-legal Solutionism: Regulating Children's Online Safety in the United States. *Proceedings of the Symposium on Computer Science and Law*, 86–97. https://doi.org/10.1145/3614407.3643705 [perma.cc/6AA8-L38R]

ARICA. (2023). *About*. ARICA. https://www.aricaproject.eu/about/ [perma.cc/DD34-5HM2]

Baines, V. (2019). Online child sexual exploitation: Towards an optimal international response. *Journal of Cyber Policy, 4*(2), 197–215. https://doi.org/10.1080/23738871.2019.1635178 [https://perma.cc/P6S9-SDM9]

Bhatia, A. (2024, September 11). The Future of the Christchurch Call Foundation and Lessons for Multistakeholder Initiatives. *Center for Democracy and Technology*. https://cdt.org/insights/the-future-of-the-christchurch-call-foundation-and-lessons-for-multistakeholder-initiatives/ [perma.cc/Q3JV-CQ54]

Bhatia, A., & Aboulafia, A. (2024, September 24). *Age Verification Technology Would Create New Barriers for Young Disabled People*. Teen Vogue. https://www.teenvogue.com/story/age-verification-technology-disabled-people [perma.cc/XP8C-ECNX]

Blake, P. (2019). Age verification for online porn: More harm than good? *Porn Studies*, 6(2), 228–237. https://doi.org/10.1080/23268743.2018.1555054 [https://perma.cc/86KF-LBC3]

Boburg, S., Verma, P., & Dehghanpoor, C. (2024, March 13). On popular online platforms, predatory groups coerce children into self-harm. *Washington Post*. https://www.washingtonpost.com/investigations/interactive/2024/764-predator-discord-telegram/ [https://perma.cc/S3ME-9UFC]

Brewer, J., Romine, M., & Taylor, T. L. (2020). Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. *Proceedings of the 2020 ACM Designing Interactive Systems Conference, 757–769*. https://doi.org/10.1145/3357236.3395514 [perma.cc/A9SB-K5K7]

Breyer, P. (2024, April 10). Pirate lawsuit: German Regional Court refuses to rule on legality of voluntary chat control scanning of private messages. *Patrick Breyer*. https://www.patrick-breyer.de/en/pirate-lawsuit-german-regional-court-refuses-to-rule-on-legality-of-voluntary-chat-control-scanning-of-private-messages/ [perma.cc/NB22-AN2Q]

Brunton, F. (2013). Spam: *A shadow history of the Internet*. MIT Press. https://doi.org/10.7551/mitpress/9384.001.0001 [https://perma.cc/Q7VQ-AWUJ]

Cai, J., Chowdhury, S., Zhou, H., & Wohn, D. Y. (2023). Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 1–28. https://doi.org/10.1145/3610191 [perma.cc/D2XG-4J8C]

Cai, J., & Wohn, D. Y. (2019). Categorizing Live Streaming Moderation Tools: An Analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies (IJICST)*, 9(2), 36–50. https://doi.org/10.4018/IJICST.2019070103 [perma.cc/T6ET-9EU7]

Cai, J., & Wohn, D. Y. (2021). After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 410:1-410:25. https://doi.org/10.1145/3479554 [perma.cc/CT9K-RHF6]

Caplan, R. (2023). Networked Platform Governance: The Construction of the Democratic Platform. *International Journal of Communication*, 17(22). Retrieved from https://ijoc.org/index.php/ijoc/article/view/20035 [perma.cc/RWX7-DNQX]

Celiksoy, E., Schwarz, K., & Sawyer, L. (2023). *Legal and institutional responses to the online sexual exploitation of children* |. University of Nottingham Rights Lab. https://www.nottingham.ac.uk/research/beacons-of-excellence/rights-lab/resources/reports-and-briefings/2023/october/legal-and-institutional-responses-to-the-online-sexual-exploitation-of-children-the-philippines-country-case-study.pdf [perma.cc/8DXY-3C8C]

Child Rights International Network & defenddigitalme. (2023). *Privacy and Protection: A children's rights approach to encryption*. Child Rights International Network and defenddigitalme. https://home.crin.org/readlistenwatch/stories/privacy-and-protection [perma.cc/Y888-H7X8]

Christensen, L. S., & Woods, J. (2024). "It's Like POOF and It's Gone": The Live-Streaming of Child Sexual Abuse. *Sexuality & Culture, Online First*. https://doi.org/10.1007/s12119-023-10186-9 [perma.cc/35L8-HGLG]

Cloudflare. (n.d.). *What is live streaming? | How live streaming works*. Retrieved October 14, 2024, from https://www.cloudflare.com/learning/video/what-is-live-streaming/ [https://perma.cc/QN58-XKGW]

CNIL. (2022). *Online age verification: Balancing privacy and the protection of minors*. https://www.cnil.fr/en/online-age-verification-balancing-privacy-and-protection-minors [perma.cc/Z4H6-4BPJ]

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology, 34*(4), 739–766. https://doi.org/10.1007/s13347-020-00429-0 [perma.cc/U5RF-R2Y2]

Cooper, K., Quayle, E., Jonsson, L., & Svedin, C. G. (2016). Adolescents and self-taken sexual images: A review of the literature. *Computers in Human Behavior*, 55, 706–716. https://doi.org/10.1016/j.chb.2015.10.003 [perma.cc/83XT-59WL]

Cox, J. (2024a, March 28). Criminals Are Weaponizing Child Abuse Imagery to Ban Discord Servers. *404 Media*. https://www.404media.co/criminals-are-weaponizing-child-abuse-imagery-to-ban-discord-servers/ [perma.cc/S79A-V5PD]

Cox, J. (2024b, June 26). *ID Verification Service for TikTok, Uber, X Exposed Driver Licenses*. 404 Media. https://www.404media.co/id-verification-service-for-tiktok-uber-x-exposed-driver-licenses-au10tix/ [perma.cc/G4NZ-49Q9]

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410–428. https://doi.org/10.1177/1461444814543163 [https://perma.cc/9HMD-FB2Z]

Crisan, A., Drouhard, M., Vig, J., & Rajani, N. (2022). Interactive Model Cards: A Human-Centered Approach to Model Documentation. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439. https://doi.org/10.1145/3531146.3533108 [perma.cc/MP8M-QL7M]

D'Anastasio, C. (2022, September 21). Child Predators Use Amazon's Twitch to Systematically Track Kids Who Stream. *Bloomberg*. https://www.bloomberg.com/graphics/2022-twitch-problem-with-child-predators/?sref=P6Q0mxvj [perma.cc/BZJ9-UGAX]

D'Anastasio, C. (2024, January 5). Twitch "Clips" Feature Being Used to Exploit Minors. *Bloomberg*. https://www.bloomberg.com/news/articles/2024-01-05/twitch-clips-feature-being-used-to-exploit-minors [perma.cc/8EAK-832R]

Denyer Willis, G. (2023). 'Trust and safety': Exchange, protection and the digital market–fortress in platform capitalism. Socio-Economic Review, 21(4), 1877–1895. https://doi.org/10.1093/ser/mwad003 [perma.cc/FVD4-VTYJ]

Drejer, C., Riegler, M. A., Halvorsen, P., Johnson, M. S., & Baugerud, G. A. (2024). Livestreaming technology and online child sexual exploitation and abuse: A scoping review. *Trauma, Violence, & Abuse*, 25(1), 260–274. https://doi.org/10.1177/15248380221147564 [https://perma.cc/A4VH-NTPG]

Drejer, C., Sabet, S. S., Baugerud, G. A., & Riegler, M. A. (2024). *It's All in the Game—An Exploration of Extensive Communication on Gaming Platforms and the Risks of Online Sexual Grooming* (SSRN Scholarly Paper 4671140). https://doi.org/10.2139/ssrn.4671140 [https://perma.cc/7UQT-EDPT]

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy & Technology. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/ [perma.cc/9DES-3EFJ]

EDRi. (2023, August 29). *Is this the most criticised draft EU law of all time?* European Digital Rights (EDRi). https://edri.org/our-work/most-criticised-eu-law-of-all-time/ [perma.cc/4MAJ-KN8M]

Europol. (2024, July 2). *Operational sprint generates 197 new leads on buyers of 'live distant child abuse.'* Europol. https://www.europol.europa.eu/media-press/newsroom/news/operational-sprint-generates-197-new-leads-buyers-of-live-distant-child-abuse [perma.cc/ETE3-HGMG]

Farid, H. (2022). Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety, 1*(4), Article 4. https://doi.org/10.54501/jots.v1i4.56 [perma.cc/X83Y-L9ZJ]

Forland, S., Meysenburg, N., & Solis, E. (2024). *Age Verification: The Complicated Effort to Protect Youth Online*. Open Technology Institute. http://newamerica.org/oti/reports/age-verification-the-complicated-effort-to-protect-youth-online/ [perma.cc/FRE2-NFJ4]

Goggin, B. (2023, June 21). Discord servers used in child abductions, crime rings, sextortion. *NBC News*. https://www.nbcnews.com/tech/social-media/discord-child-safety-social-platform-challenges-rcna89769 [perma.cc/SG7P-Q3QC]

Google. (2024a). *Create a live stream on mobile*. https://support.google.com/youtube/answer/9228390 [perma. cc/LKY4-2VNE]

Google. (2024b). *Verify your YouTube account—YouTube Help*. https://support.google.com/youtube/ answer/171664?hl=en [perma.cc/3VSR-DX9B]

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 205395171989794. https:// doi.org/10.1177/2053951719897945 [https://perma.cc/9RWX-PKM6]

Gorwa, R., & Veale, M. (forthcoming). *Routine Resistant Platform Governance* (Working Paper).

Grossman, S., Pfefferkorn, R., Thiel, D., Shah, S., Stamos, A., DiResta, R., Perrino, J., Cryst, E., & Hancock, J. (2024). *The Strengths and Weaknesses of the Online Child Safety Ecosystem: Perspectives from Platforms, NCMEC, and Law Enforcement on the CyberTipline and How to Improve It*. https://doi.org/10.25740/ pr592kc5483 [perma.cc/GA64-7LEY]

Han, C., Seering, J., Kumar, D., Hancock, J. T., & Durumeric, Z. (2023). Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–28. https://doi.org/10.1145/3579609 [perma.cc/5L6B-FLVF]

Horsman, G. (2018). A forensic examination of the technical and legal challenges surrounding the investigation of child abuse on live streaming platforms: A case study on Periscope. *Journal of Information Security and Applications, 42*, 107–117. https://doi.org/10.1016/j.jisa.2018.07.009 [perma.cc/S5J9-W5EE]

Insoll, T., Ovaska, A., & Vaaranen-Valkonen, N. (2021). *CSAM Users in the Dark Web: Protecting Children Through Prevention*. Suojellaan Lapsia/Protect Children. https://www.suojellaanlapsia.fi/en/post/csam-users-in-the-dark-web-protecting-children-through-prevention [perma.cc/JM6G-ALNN]

International Justice Mission & University of Nottingham Rights Lab. (2023). *Scale of Harm: Estimating the Prevalence of Trafficking to Produce Child Sexual Exploitation Material in the Philippines*. International Justice Mission. https://www.ijm.org/studies/scale-of-harm-estimating-the-prevalence-of-trafficking-to-produce-child-sexual-exploitation-material-in-the-philippines [perma.cc/8JB9-2JXB]

Jackson, G. (2019, October 14). Twitch Streamer Says She Was Banned For "Suggestive" Attire After Brigade From Racist Trolls. *Kotaku*. https://kotaku.com/twitch-streamer-says-she-was-banned-for-suggestive-atti-1839040894 [perma.cc/GQ7K-KP96]

Kamara, S., Knodel, M., Llansó, E., Nojeim, G., Qin, L., Thakur, D., & Vogus, C. (2021). *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems* (p. 38). Center for Democracy & Technology. https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/ [perma.cc/97V3-M8H2]

Kennedy, Ü., Lala, G., Rajan, P., Sardarabady, S., & Tatam, L. (2024). *Protecting Children from Online Grooming: Cross-cultural, qualitative and child-centred data to guide grooming prevention and response*. Save the Children. https://resourcecentre.savethechildren.net/document/protecting-children-from-online-grooming-cross-cultural-qualitative-and-child-centred-data-to-guide-grooming-prevention-and-response/ [https://perma.cc/ED6M-LM4T]

Laranjeira da Silva, C., Macedo, J., Avila, S., & dos Santos, J. (2022). Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2189–2205. https://doi.org/10.1145/3531146.3534636 [perma.cc/2MLT-6Z9V]

Llansó, E. (2020, July 30). Human Rights NGOs in Coalition Letter to GIFCT. *Center for Democracy and Technology*. https://cdt.org/insights/human-rights-ngos-in-coalition-letter-to-gifct/ [perma.cc/NZZ6-XDPU]

Luria, M. (2023). *More Tools, More Control: Lessons from Young Users on Handling Unwanted Messages Online*. Center for Democracy & Technology. https://cdt.org/insights/more-tools-more-control-lessons-from-young-users-on-handling-unwanted-messages-online/ [perma.cc/L756-HP44]

Marwick, A., Smith, J., Caplan, R., & Wadhawan, M. (2024). Child Online Safety Legislation (COSL)—A Primer. *The Bulletin of Technology & Public Life*. https://doi.org/10.21428/bfcb0bff.de78f444 [perma.cc/A5LE-YVRL]

McAlinden, A.-M. (2006). 'Setting 'Em Up': Personal, Familial and Institutional Grooming in the Sexual Abuse of Children. *Social & Legal Studies*, 15(3), 339–362. https://doi.org/10.1177/0964663906066613 [https://perma.cc/YBW5-M7RW]

McKee, A., & Lumby, C. (2022). Pornhub, child sexual abuse materials and anti-pornography campaigning. *Porn Studies*, 9(4), 464–476. https://doi.org/10.1080/23268743.2022.2083662 [https://perma.cc/Q7ND-7FWN]

Meta. (2023, December 1). Our Work To Fight Online Predators. *Meta*. https://about.fb.com/news/2023/12/combating-online-predators/ [perma.cc/MBL8-9DUJ]

Meta. (2024a). *Child Sexual Exploitation, Abuse, and Nudity | Transparency Center*. https://transparency.meta.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/ [https://perma.cc/A4ES-MERB]

Meta. (2024b). *IndicatorType—ThreatExchange—Documentation*. Meta for Developers. https://developers.facebook.com/docs/threat-exchange/reference/apis/indicator-type/v21.0/ [https://perma.cc/8RM5-T35N]

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596 [perma.cc/VEQ7-K2D8]

MSAB. (2023, May 31). *Safer Digital Spaces: The Vital Role of Technology in Combating CSAM*. MSAB. https://www.msab.com/blog/forensic-fix-tom-farrell-jesse-nicholson/ [perma.cc/DT3U-4UPE]

Nicholas, G., & Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis*. Center for Democracy & Technology. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ [perma.cc/Y7JL-F5GW]

Payt, S. (2024, September 27). *Council Post: 3 Solutions To The Technology-Facilitated Crimes Against Children*. Forbes. https://www.forbes.com/councils/forbesnonprofitcouncil/2024/09/27/3-solutions-to-the-technology-facilitated-crimes-against-children/ [perma.cc/B8Q8-535U]

Peralta, D. (2023). AI and suicide risk prediction: Facebook live and its aftermath. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01651-y [perma.cc/C3Z7-YNNN]

Pereira, M., Dodhia, R., Anderson, H., & Brown, R. (2023). Metadata-Based Detection of Child Sexual Abuse Material. IEEE Transactions on Dependable and Secure Computing, 1–13. *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/TDSC.2023.3324275 [perma.cc/HAJ3-ZAXK]

Persson, J. (2024). Age as a Gatekeeper in the UK Online Safety Agenda. In E. Setty, F. Gordon, & E. Nottingham (Eds.), *Children, Young People and Online Harms: Conceptualisations, Experiences and Responses* (pp. 169–181). Springer International Publishing. https://doi.org/10.1007/978-3-031-46053-1_7 [perma.cc/EYF9-G85L]

Pfefferkorn, R. (2023, December 19). *Child Safety-Focused REPORT Act Passes US Senate | TechPolicy.Press*. Tech Policy Press. https://techpolicy.press/child-safetyfocused-report-act-passes-us-senate [perma.cc/QCA5-EM4K]

Quayle, E. (2020). Prevention, disruption and deterrence of online child sexual exploitation and abuse. *ERA Forum*, 21(3), 429–447. https://doi.org/10.1007/s12027-020-00625-7 [perma.cc/7PDJ-L5U3]

Quayle, E. (2022). Self-produced images, sexting, coercion and children's rights. ERA Forum, 23(2), 237–251. https://doi.org/10.1007/s12027-022-00714-9 [perma.cc/PBE3-RVFD]

Reyes, I., Wijesekera, P., Reardon, J., Elazari Bar On, A., Razaghpanah, A., Vallina-Rodriguez, N., & Egelman, S. (2018, July 24). *"Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale*. The 18th Privacy Enhancing Technologies Symposium (PETS 2018). https://dspace.networks.imdea.org/handle/20.500.12761/551 [perma.cc/YNW9-8CHG]

Ruane, K., Branum, B., Doty, N., & Jain, S. (2024, September 23). CDT Files Amicus Brief in Free Speech Coalition v. Paxton, Challenging TX Age Verification Law. *Center for Democracy and Technology*. https://cdt.org/insights/cdt-files-amicus-brief-in-free-speech-coalition-v-paxton-challenging-tx-age-verification-law/ [perma.cc/YG2W-263P]

Ruberg, B. (2021). "Obscene, pornographic, or otherwise objectionable": Biased definitions of sexual content in video game live streaming. *New Media & Society*, 23(6), 1681–1699. https://doi.org/10.1177/1461444820920759 [https://perma.cc/4NZP-CB8N]

Salter, M., & Sokolov, S. (2024). "Talk to strangers!" Omegle and the political economy of technology-facilitated child sexual exploitation. *Journal of Criminology*, 57(1), 121–137. https://doi.org/10.1177/26338076231194451 [https://perma.cc/SV28-5TWN]

Setter, C., Greene, N., Newman, N., & Perry, J. (2021). *Global Threat Assessment 2021*. WeProtect Global Alliance. https://www.weprotect.org/global-threat-assessment-21/#report [perma.cc/ZKY9-9VKK]

Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*. Center for Democracy and Technology. https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/ [perma.cc/5H78-DF8K]

Stardust, Z., Obeid, A., McKee, A., & Angus, D. (2024). Mandatory age verification for pornography access: Why it can't and won't 'save the children.' *Big Data & Society*, 11(2), 20539517241252129. https://doi.org/10.1177/20539517241252129 [https://perma.cc/EB9M-F7PX]

Stripchat. (2024, February 28). *What documents do I need to upload to create my account?* Stripchat FAQ. https://support.stripchat.com/hc/en-us/articles/4410734320785-What-documents-do-I-need-to-upload-to-create-my-account [https://perma.cc/ADZ4-ML7G]

Tech Coalition. (2022). *Tech Coalition | Trust: Voluntary Framework for Industry Transparency*. Tech Coalition. https://www.technologycoalition.org/knowledge-hub/trust-voluntary-framework-for-industry-transparency [perma.cc/ER2S-DH6U]

Tech Coalition. (2023, November 7). *Tech Coalition | Announcing Lantern: The First Child Safety Cross-Platform Signal Sharing Program*. Tech Coalition. https://www.technologycoalition.org/newsroom/announcing-lantern [perma.cc/NV2D-2PPW]

Teunissen, C., & Napier, S. (2023). The overlap between child sexual abuse live streaming, contact abuse and other forms of child exploitation. *Trends and Issues in Crime and Criminal Justice, 671*, 1–16. https://www.aic.gov.au/publications/tandi/tandi671 [https://perma.cc/34HA-8NQW]

Teunissen, C., Napier, S., & Boxall, H. (2021). Live streaming of child sexual abuse: An analysis of offender chat logs. *Trends and Issues in Crime and Criminal Justice, 639*, 1–15. https://doi.org/10.52922/ti78375 [https://perma.cc/J5AN-FM9T]

Thiel, D., DiResta, R., & Stamos, A. (2023). *Cross-Platform Dynamics of Self-Generated CSAM*. Stanford Internet Observatory. https://purl.stanford.edu/jd797tp7663 [perma.cc/CH99-A4YB]

Thorn. (n.d.). *Text Classifier for Child Safety | Safer Predict, Built by Thorn*. Retrieved October 15, 2024, from https://get.safer.io/text-classification-content-moderation [perma.cc/UVG7-YZ99]

Thorn. (2022, September 23). *How CSAM Detection Works | Safer by Thorn*. Safer: Proactive Solution for CSE and CSAM Detection. https://safer.io/how-it-works/ [perma.cc/Z5F9-3RDK]

Thorn. (2024, June 26). *CSAM Keyword Hub Application | Safer.i*o. Safer: Proactive Solution for CSE and CSAM Detection. https://safer.io/resources/csam-keyword-hub/ [perma.cc/5TUY-G7XU]

TikTok. (2024a). *Minimum age appeals on TikTok | TikTok Help Center*. https://support.tiktok.com/en/safety-hc/account-and-user-safety/minimum-age-appeals-on-tiktok [perma.cc/G84K-CNXB]

TikTok. (2024b). *Protecting teens online*. https://www.tiktok.com/transparency/en-us/protecting-teens/ [perma.cc/5TCH-PKWK]

TikTok. (2024c, January 19). *LIVE Center*. https://livecenter.tiktok.com/help_center/article/1023/tiktok-live-studio-operation-manual_en-US?lang=en [perma.cc/QQ8K-UMLF]

TikTok. (2024d, September 26). *Community Guidelines Enforcement Report—April 1—June 30, 2024*. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2024-9 [perma.cc/U6NN-KVKP]

TikTok. (2024e, October 25). *TikTok Creator Academy: Empowering Creators to Grow and Succeed on TikTok | TikTok For Creator*. https://www.tiktok.com/creator-academy/en/article/Going-LIVE?ref=kapwing-resources [perma.cc/45E2-2CF5]

Tirfe, D., & Anand, V. K. (2022). A Survey on Trends of Two-Factor Authentication. In H. K. D. Sarma, V. E. Balas, B. Bhuyan, & N. Dutta (Eds.), *Contemporary Issues in Communication, Cloud and Big Data Analytics* (pp. 285–296). Springer. https://doi.org/10.1007/978-981-16-4244-9_23 [perma.cc/UGF8-ALSA]

Tommy I. (2023, January 10). The Subscriber Requirements For Livestreaming On YouTube: How To Get Started | *TuBeast.com*. | TuBeast.Com. https://tubeast.com/the-subscriber-requirements-for-livestreaming-on-youtube-how-to-get-started [perma.cc/E8MF-QNUP]

Twitch. (n.d.). *Chat Verification Settings*. Retrieved October 14, 2024, from https://help.twitch.tv/s/article/chat-verification-settings?language=en_US [perma.cc/VUS7-LGN3]

Twitch. (2022, November 22). *Our Ongoing Work to Combat Online Grooming*. https://safety.twitch.tv/s/article/Our-Work-to-Combat-Online-Grooming?language=en_US [perma.cc/XU6Y-F9TJ]

Twitch. (2023). *H1 2023 Transparency Report*. Twitch. https://safety.twitch.tv/s/article/H1-2023-Transparency-Report?language=en_US [perma.cc/D7NR-C7UQ]

Twitch. (2024, March 26). *Twitch.tv—Terms of Service.* Twitch.Tv. https://www.twitch.tv/p/en/legal/terms-of-service/#2-use-of-twitch-by-minors-and-blocked-persons [perma.cc/8V7P-6YLU]

Vallance, C. (2024, April 22). *Three-year-olds groomed online, Internet Watch Foundation warns*. https://www.bbc.com/news/articles/cx9wezr1d1vo [perma.cc/6ZRW-3SD5]

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 336–349. https://doi.org/10.1145/3531146.3533101 [perma.cc/G6E2-FGDD]

Winslow, L. (2024, January 5). Report: Predators Are Using Twitch "Clips" To Spread Child Abuse. *Kotaku*. https://kotaku.com/twitch-clips-feature-predators-child-abuse-tiktok-1851144631 [perma.cc/U56B-MQ5T]

Witting, S. K. (2019). Regulating bodies: The moral panic of child sexuality in the digital era. *Kritische Vierteljahresschrift Für Gesetzgebung Und Rechtswissenschaft, 102*(1), 5–38. https://doi.org/10.5771/2193-7869-2019-1-5 [perma.cc/67PT-WV94]

Xiao, F. (2024). Moderating for a friend of mine: Content moderation as affective reproduction in Chinese live-streaming. *Media, Culture & Society, 46*(1), 60–77. https://doi.org/10.1177/01634437231188465 [https://perma.cc/25KW-9SAH]

Zhao, D., Wang, A., & Russakovsky, O. (2021). Understanding and Evaluating Racial Biases in Image Captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14830–14840. https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Understanding_and_Evaluating_Racial_Biases_in_Image_Captioning_ICCV_2021_paper.html [perma.cc/8ZXT-CC2T]

Zornetta, A., & Pohland, I. (2022). Legal and technical trade-offs in the content moderation of terrorist live-streaming. *International Journal of Law and Information Technology, 30*(3), 302–320. https://doi.org/10.1093/ijlit/eaac020 [perma.cc/Y2CV-YJTJ]

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY