



Amenazas en Tiempo Real:

**Análisis de Prácticas de Confianza y Seguridad para la
Prevención de la Explotación y el Abuso Sexual Infantil en
Plataformas de Transmisión en Vivo**

**Robert Gorwa
Dhanaraj Thakur**

Diciembre 2024



The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

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.

Amenazas en Tiempo Real

Análisis de Prácticas de Confianza y Seguridad para la Prevención de la Explotación y el Abuso Sexual Infantil en Plataformas de Transmisión en Vivo

Autores: Robert Gorwa y Dhanaraj Thakur*

CON CONTRIBUCIONES DE

Samir Jain, Gabriel Nicholas, Aliya Bhatia, Kate Ruane, Mallory Knodel, Silvia Lorenzo Perez, y Drew Courtney.

AGRADECIMIENTOS

Agradecemos a Riana Pfefferkorn, Jen Persson y David Thiel por su revisión y comentarios sobre los primeros borradores de este informe. También agradecemos a todos los participantes en nuestro taller de junio de 2024 que brindaron comentarios sobre nuestros hallazgos iniciales, así como también agradecemos a quienes ayudaron a implementarlo, incluyendo a Jamal Magby, DeVan L. Hankerson, Ozzie Oguine, Noor Waheed y Saanvi Arora. Por último, gracias a todas las personas que entrevistamos y que tuvieron la amabilidad de compartir su experiencia y sus conocimientos sobre el tema de este informe. Todas las conclusiones y recomendaciones formuladas en este informe son las del CDT.

Diseño de portada y maquetación: Gabriel Hongsdusit.

Dirección de arte: Timothy Hoagland.

El Centro para la Democracia y la Tecnología (CDT) agradece el apoyo financiero brindado para este proyecto por Safe Online. Esta publicación ha sido elaborada con el apoyo financiero de Safe Online. Sin embargo, las opiniones, hallazgos, conclusiones y recomendaciones expresadas en este documento son las de CDT y no reflejan necesariamente las de Safe Online.



Cita sugerida: Gorwa, R. y Thakur, D. (2024) *Amenazas en Tiempo Real: Análisis de Prácticas de Confianza y Seguridad para la Prevención de la Explotación y el Abuso Sexual Infantil en Plataformas de Transmisión en Vivo*. Centro para la Democracia y la Tecnología. <https://cdt.org/insights/real-time-threats-analysis-of-trust-and-safety-practices-for-child-sexual-exploitation-and-abuse-csea-prevention-on-livestreaming-platforms/>

* Autor correspondiente: research@cdt.org

Las referencias en este informe incluyen enlaces originales, así como enlaces archivados y acortados por el servicio Perma.cc. Los enlaces de Perma.cc también contienen información sobre la fecha de recuperación y archivo.



Índice

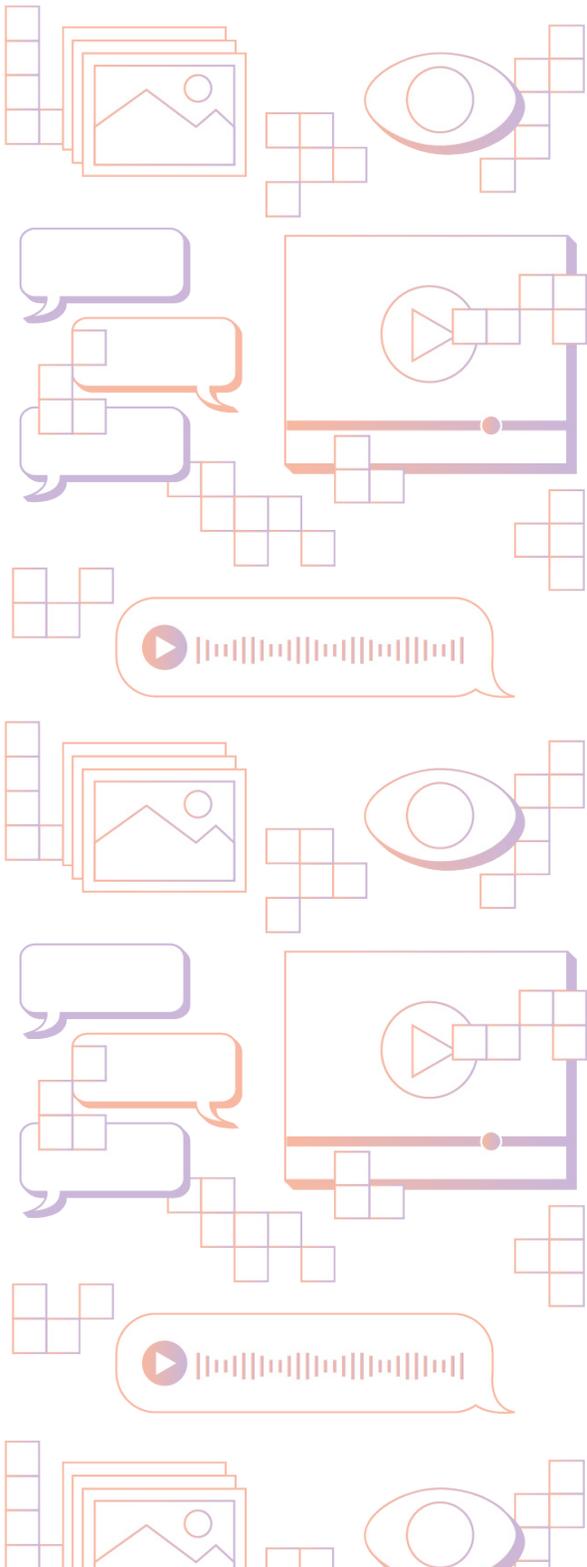
Resumen Ejecutivo	5
Introducción	9
Investigación sobre la Transmisión en vivo y la Explotación y el Abuso Sexual Infantil	12
Alcance y Métodos de Denuncia	15
I. Protección de las Transmisiones: Una Visión General de la Confianza y la Seguridad	18
Diseño: Herramientas de Verificación, Autenticación y Moderación de la Comunidad	22
Análisis de Contenido: Escrutinio de las Transmisiones en busca de Violaciones a las Políticas	25
Señales: Investigación, Seguimiento y Predicción de Comportamientos Violatorios	30
II. Implicaciones Políticas de las Prácticas de Confianza y Seguridad Existentes	34
Limitaciones de las Intervenciones de Diseño Existentes	34
Ramificaciones del Análisis Predictivo de Vídeos en Transmisiones En Vivo: Sesgo, Sobrebloqueo, Calidad de Datos y Fuentes de Datos	35
Problemas con el uso compartido de señales: Privacidad, Opacidad y Pocos Recursos	38
Conclusión	41
Referencias	45

Resumen Ejecutivo

En los últimos años han surgido una serie de nuevos servicios en línea que facilitan la transmisión en vivo (livestreaming) de video y audio en tiempo real. A través de estas herramientas, los usuarios y creadores de contenido de todo el mundo pueden transmitir fácilmente sus actividades a audiencias globales potencialmente grandes, facilitando formas participativas, generadoras y colaborativas “en vivo” de juegos, creación de música, debates y otras interacciones. Sin embargo, el ascenso de estas plataformas no ha sido perfecto. Estas mismas herramientas se utilizan para difundir contenido socialmente problemático y/o ilegal, desde la promoción de la autolesión y el extremismo violento hasta materiales para la explotación y el abuso sexual infantil (CSEA, por sus siglas en inglés).

Este informe examina la gama de herramientas y prácticas de confianza y seguridad que están desarrollando e implementando las plataformas y los proveedores externos para proteger los servicios de transmisión en vivo, con un enfoque especial en la prevención de la explotación y el abuso sexual infantil. Moderar los medios en tiempo real es inherentemente difícil desde el punto de vista técnico para las empresas que buscan intervenir de manera responsable, ya que gran parte del contenido transmitido en vivo es “nuevo”, producido en el momento y, por lo tanto, por definición, no es “conocido” y no es posible compararlo con material dañino previamente identificado mediante técnicas basadas en hash (huella digital). Las empresas que buscan analizar transmisiones en vivo deben hacerlo con modelos de visión artificial predictivos comparativamente ineficientes y potencialmente defectuosos, trabajando creativamente con el audio de la transmisión (por ejemplo, mediante la transcripción y la clasificación de texto) y/o mediante otras técnicas emergentes, como intervenciones orientadas a “señales” basadas en las características de comportamiento de cuentas de usuarios sospechosas.

Con base en una revisión de los documentos disponibles públicamente sobre plataformas de transmisión en vivo y de proveedores que ofrecen servicios de análisis de contenido, así como en entrevistas con personas que trabajan en este problema en la industria, la sociedad civil y con académicos, encontramos que la industria está adoptando tres enfoques principales para abordar la explotación y el abuso sexual infantil en la transmisión en vivo:



- **Enfoques basados en el diseño:** Pasos que se toman antes de que un usuario pueda transmitir, como implementar medidas de fricción y verificación destinadas a dificultar que los usuarios, o los usuarios sospechosos, puedan transmitir en vivo. Por ejemplo, algunas plataformas exigen que un usuario tenga un número determinado de seguidores o suscriptores antes de poder transmitir en vivo para evitar que un alguien cree espontáneamente una cuenta y transmita en vivo contenido dañino.
- **Enfoques de análisis de contenido:** Diversas formas de detección y análisis de contenido manual o automatizado que pueden funcionar en video, audio y texto mientras el contenido se transmite en vivo. Los ejemplos incluyen tomar fotogramas de muestra de transmisiones en vivo y ver si coinciden con hashes de material ya conocido de explotación y abuso sexual infantil; uso de clasificadores de aprendizaje automático para detectar material de abuso sexual infantil (CSAM, por sus siglas en inglés) en video en vivo; y empleo análisis predictivo de transcripciones de texto de audio en vivo o chats de usuarios en transmisiones en vivo.
- **Enfoques basados en señales:** Intervenciones basadas en las características conductuales y los metadatos de las cuentas de usuario. Por ejemplo, las plataformas pueden compartir ciertos metadatos de cuentas para ayudar a identificar actores maliciosos a medida que se mueven de una plataforma a otra o usar señales para identificar cuentas involucradas en un comportamiento potencialmente sospechoso que impulse a una mayor investigación.

En parte debido a los desafíos que plantea la detección de contenido en vivo, está evolucionando la forma en que la industria aborda el problema de la explotación y el abuso sexual infantil, y otros contenidos dañinos. Como lo expresó un entrevistado, la idea es que las empresas participen más activamente en la reducción de la capacidad de usar su plataforma para la propagación de la explotación y el abuso sexual infantil, no solo participando en un modo de detección y denuncia, sino también, aspiracionalmente, hacia un modelo de predicción e interrupción de confianza y seguridad más similar al utilizado en áreas como la ciberseguridad y el fraude.

Los enfoques de la industria respecto a la explotación y el abuso sexual infantil plantean varias preocupaciones. En primer lugar, existe una tendencia general a evitar la transparencia y la claridad respecto a la forma en que estos sistemas funcionan y se implementan, aparentemente para evitar que actores maliciosos los eludan, pero en última instancia en detrimento de las víctimas, los usuarios, los responsables políticos y otras partes interesadas. En segundo lugar, y relacionado con el primer punto, es casi imposible determinar cuán eficaces son estos enfoques, qué vacíos dejan, si resultan en una moderación excesiva de contenido legítimo y cuán bien atienden las necesidades de todas las partes interesadas. En tercer lugar, estos enfoques introducen importantes riesgos para la seguridad, privacidad, libertad de expresión y otros derechos humanos que pueden socavar la seguridad de los menores a quienes pretenden proteger, así como la de los usuarios en general.

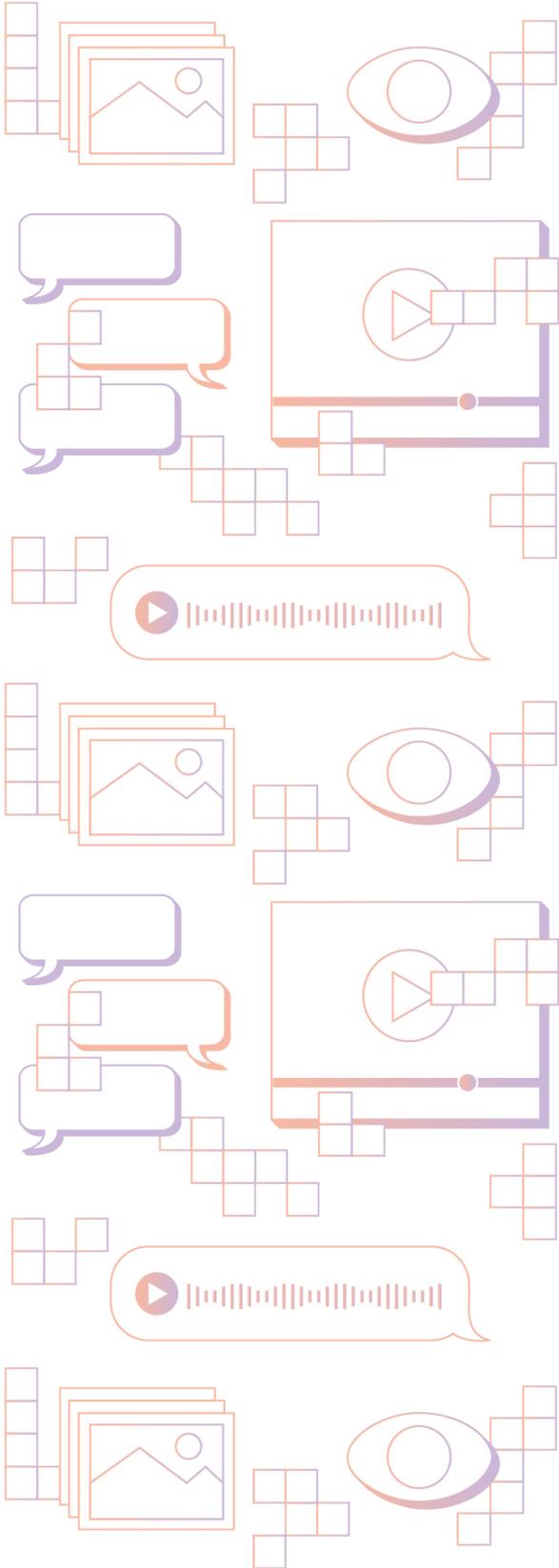
Para ayudar a abordar estas preocupaciones, destacamos cuatro áreas de mejora:

1. **Se necesita una mayor transparencia para ayudar a evaluar y mejorar los esfuerzos para abordar la explotación y el abuso sexual infantil en plataformas de transmisión en vivo.** Por ejemplo, actualmente no existen métricas de desempeño que las empresas puedan usar para probar y comparar la precisión de las medidas que toman o que los expertos, los responsables políticos y los investigadores puedan usar para comprender mejor su eficacia, así como el alcance de lo que realmente es posible.
2. **Los proveedores y las plataformas de transmisión en vivo deben ser explícitos acerca de las limitaciones de los enfoques automatizados para detectar y abordar la explotación y el abuso sexual infantil.** Al hacerlo, las plataformas pueden mejorar sus sistemas de confianza y seguridad al garantizar que los revisores humanos participen adecuadamente y permitirles tomar decisiones llenas de matices basadas en el contexto y otra información.
3. **Centrarse en intervenciones de diseño que empoderen a los usuarios, incluidos los menores.** Las necesidades de los streamers de protegerse de ser blanco de ataques o de ser utilizados para distribuir la explotación y el abuso sexual infantil merecen mayor atención cuando se trata de soluciones basadas en el diseño. Por ejemplo, un enfoque basado en el diseño que no se planteó en nuestras conversaciones con la industria es proporcionar a los usuarios, en particular a los menores, el conjunto adecuado de herramientas y mecanismos de denuncia para ayudarles a protegerse a sí mismos y a los demás.
4. **Los modelos de gobernanza de múltiples partes interesadas pueden mejorar la rendición de cuentas de los enfoques para abordar la explotación y el abuso sexual infantil en las transmisiones en vivo.** Los marcos de mejores prácticas en torno a la implementación de estos sistemas podrían desarrollarse no solo a través del trabajo continuo de organizaciones como la Tech Coalition, sino también a través de la participación crítica de múltiples partes interesadas en foros que no solo involucren a organizaciones de seguridad infantil, sino también a organizaciones que participan activamente en un conjunto más amplio de derechos digitales y libertades civiles.

Abordar el problema de la explotación y el abuso sexual infantil en general y en las plataformas de transmisión en vivo es de vital importancia dado el impacto en los niños, los padres y sus comunidades, por lo que esta es un área de gobernanza de las plataformas de enorme importancia y de alto riesgo. Es comprensible que tanto los proveedores como la industria estén ansiosos por demostrar que están desarrollando herramientas innovadoras para abordar la explotación y el abuso sexual infantil, y otro contenido dañino, pero una implementación deficiente (o un diseño deficiente, con sistemas que son fundamentalmente defectuosos) disminuirá, en lugar de aumentar, la fiabilidad de los responsables políticos y del público en la confianza y la seguridad de las plataformas en el largo plazo. Una mejor comprensión de las medidas que están adoptando las plataformas en las plataformas de transmisión en vivo, junto con una mayor participación de múltiples partes interesadas, mejorará los sistemas de confianza y seguridad de manera que se minimice el riesgo de explotación y abuso sexual infantil en el contenido transmitido en vivo, al tiempo que se minimizan los impactos no deseados en los usuarios comunes.



Introducción



En los últimos años han surgido una serie de nuevos servicios en línea que permiten la transmisión de video y audio en tiempo real. A través de estas herramientas de transmisión en vivo, los usuarios y creadores de contenido pueden distribuir contenidos a medida que se crean (en lugar de finalizar el contenido antes de comenzar la distribución). Los streamers pueden transmitir sus actividades en tiempo real a grandes audiencias globales, lo que facilita formas participativas, generadoras y colaborativas “en vivo” de juegos, creación musical, debates y otras interacciones. Generalmente, esto se hace de una a muchas personas de diversas maneras, como en transmisiones públicas o en grupos privados. La mayoría de los servicios de transmisión en vivo utilizan una combinación de compresión, codificación y redes de distribución de contenido para distribuir contenido en todo el mundo (Cloudflare, sin fecha).

A medida que los servicios de transmisión en vivo se vuelven más frecuentes y populares, especialmente entre los usuarios más jóvenes, las partes interesadas del gobierno, la sociedad civil y la industria han prestado cada vez más atención a las prácticas de “confianza y seguridad” de las empresas que los operan. En particular, los organismos encargados de hacer cumplir la ley, los grupos de la sociedad civil y otros han expresado constantemente su preocupación por la posibilidad de que las plataformas de transmisión en vivo se utilicen para facilitar diversas formas de explotación y abuso sexual infantil (Setter *et al.*, 2021).

En general, se entiende que la explotación y el abuso sexual infantil incluye la producción, difusión y posesión de material/imágenes de abuso sexual infantil (o contenido que incluye actividades sexualmente explícitas con niños); grooming a niños en línea con fines sexuales (El Diccionario de la Lengua Española 2024 define “grooming” como el: “acoso sexual a menores de edad, que se basa en establecer con ellos una relación de confianza a través de medios informáticos o telemáticos, fundamentalmente en chats y redes sociales”); sextorsión; y prostitución en línea que involucra a niños (Quayle, 2020). Por lo tanto, la transmisión en vivo de la explotación y el abuso sexual infantil implica la “producción, transmisión y visualización en tiempo real de abuso sexual infantil y está relacionada con la explotación sexual a través de la prostitución, actividades sexuales y la producción de explotación y abuso sexual infantil” (Drejer, Riegler *et al.*, 2024). Aunque los datos empíricos exhaustivos son escasos, existe evidencia de que varios tipos de plataformas de transmisión en vivo se han convertido en un vector para la producción, difusión y consumo de la explotación y el abuso sexual infantil (Insoll *et al.*, 2021).

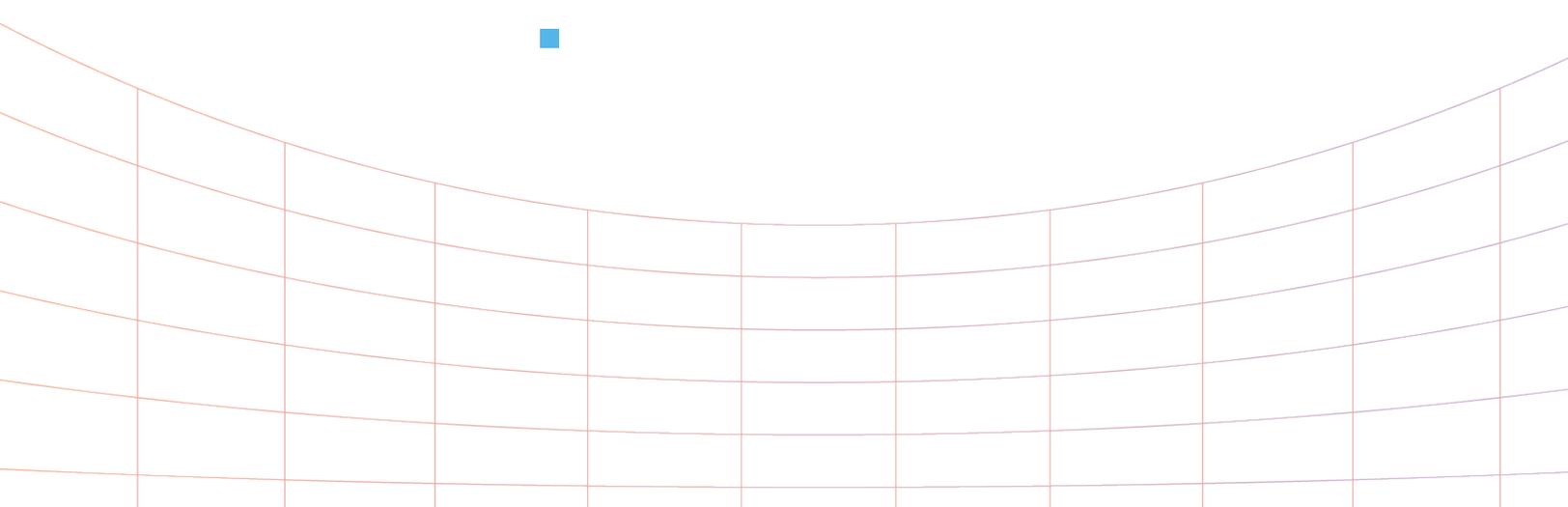
Exploramos la gama de herramientas y prácticas de confianza y seguridad que están desarrollando varios actores de plataformas para salvaguardar sus ofertas de transmisión en vivo, buscando proporcionar un mapeo inicial de un ecosistema poco examinado y generalmente opaco que, sin embargo, podría tener impactos significativos en la experiencia en línea y los derechos de muchos usuarios de Internet en todo el mundo.

En el contexto del contenido almacenado y no transmitido en vivo, los intermediarios de Internet han implementado en la última década un conjunto de mejores prácticas mínimas destinadas a contrarrestar la proliferación de imágenes de abuso sexual infantil en línea y cumplir con los regímenes legales formales en países como Estados Unidos. Parte de ese trabajo se ha realizado a través de la colaboración global de múltiples partes interesadas facilitada por organizaciones como la Alianza Global WePROTECT y el Centro Nacional para Niños Desaparecidos y Explotados (NCMEC, por sus siglas en inglés). En el centro de estos avances se encuentran las técnicas de aprendizaje automático que facilitan la comparación computacional o “huella” de imágenes confirmadas de abuso infantil. El emparejamiento es un enfoque en el que se utilizan modelos de aprendizaje automático para analizar contenido en línea. Permite a las empresas comparar las huellas del contenido sospechoso con “bases de datos hash” de material verdadero de abuso sexual infantil que se mantienen y actualizan continuamente por organizaciones de la sociedad civil vinculadas al gobierno en diversas jurisdicciones. En otras palabras, la comparación permite a las empresas y a otros identificar un contenido como idéntico o suficientemente similar a material de abuso sexual infantil que ya se conoce (Shenkman *et al.*, 2021).

Sin embargo, la transmisión en vivo plantea desafíos técnicos inherentes para las empresas que buscan intervenir responsablemente contra material de abuso sexual infantil. Gran parte del contenido transmitido en vivo es “nuevo”, producido en el momento y, por lo tanto, por definición, no es “conocido” ni puede compararse con material existente utilizando técnicas basadas en huellas digitales (hashes). Este problema fundamental elimina la herramienta más confiable para el análisis automatizado de contenido de las cajas de herramientas de confianza y seguridad de la industria (Farid, 2022; Gorwa *et al.*, 2020). Las empresas que buscan analizar videos en vivo deben hacerlo a menudo utilizando la segunda aplicación principal del aprendizaje automático para el análisis de contenidos: Modelos predictivos. Estos modelos reconocen las características o rasgos de un fragmento de contenido basándose en el aprendizaje previo de la máquina. Sin embargo, incluso cuando se aplican a imágenes estáticas, estos modelos predictivos de visión por computadora o técnicas relacionadas suelen tener fallas (Shenkman *et al.*, 2021). Esos defectos solo se magnifican en el contexto del video en vivo. Además, aplicar estas herramientas a escala para decenas de miles o incluso más transmisiones simultáneas es computacionalmente difícil incluso para actores con buenos recursos, lo que potencialmente introduce importantes problemas de latencia y calidad que podrían socavar toda la propuesta de valor de los productos orientados al streaming en vivo.

Este informe explora cómo la industria está respondiendo a estos desafíos y algunas de las implicaciones de privacidad y de otro tipo de las soluciones que han adoptado. El Centro para la Democracia y la Tecnología (CDT, por sus siglas en inglés) ha seguido de cerca el uso de las herramientas automatizadas de análisis de contenido, tanto analizando su valor potencial como sus implicaciones para los derechos humanos y la libertad de expresión (Duarte *et al.*, 2017; Shenkman *et al.*, 2021). Aquí exploramos la gama de herramientas y prácticas de confianza y seguridad que están desarrollando varios actores de plataformas para salvaguardar sus ofertas de transmisión en vivo, buscando proporcionar un mapeo inicial de un ecosistema poco examinado y generalmente opaco que, sin embargo, podría tener impactos significativos en la experiencia en línea y los derechos de muchos usuarios de Internet en todo el mundo. Además de los esfuerzos de las principales plataformas de transmisión en vivo, ha habido una proliferación de proveedores de “tecnología de seguridad” y otros grupos de terceros que impulsan una amplia gama de soluciones técnicas para detectar y abordar la explotación y el abuso sexual infantil, que también consideramos en este informe.

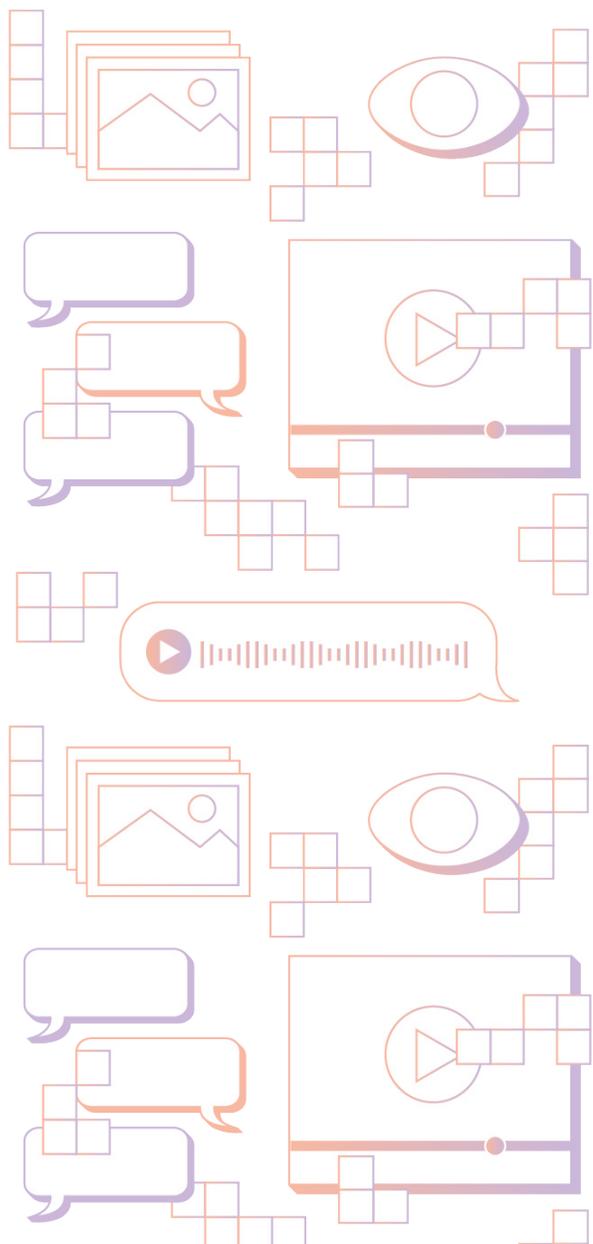
Después de una breve introducción de las investigaciones existentes, la segunda parte de este informe describe las técnicas de análisis de contenido automatizado que los servicios de transmisión en vivo utilizan actualmente para detectar la explotación y el abuso sexual infantil. Estos esfuerzos incluyen (1) intervenciones de diseño destinadas a crear fricción en el proceso de transmisión de contenido en vivo y, de ese modo, reducir el potencial de abuso, (2) varias técnicas para el análisis de transmisiones de audio y video, y (3) metadatos y medidas de creación de perfiles de comportamiento orientadas a cuentas de usuario que cada vez más se alejan de la detección de contenido problemático y se acercan a un enfoque de inteligencia de amenazas. En la tercera sección, ofrecemos un análisis de estas técnicas orientado a la formulación de políticas, incluida una evaluación de las tendencias emergentes más importantes en este ecosistema. Estos incluyen el uso cada vez mayor de perfiles de usuarios, metadatos y “señales de comportamiento” por parte de las plataformas para tratar de identificar a los actores maliciosos e intervenir contra ellos. Analizamos las posibles preocupaciones sobre transparencia, sesgo, privacidad y derechos humanos; los problemas de calidad de los datos que afectan el uso de estos sistemas; y algunas formas en que las plataformas están (y podrían ser) mejores en la implementación de la supervisión de sus esfuerzos emergentes para minimizar los impactos no deseados en los usuarios comunes.



I. Investigación sobre la Transmisión en vivo y la Explotación y el Abuso Sexual Infantil

La investigación sobre el abuso sexual infantil y la transmisión en vivo es difícil de navegar, ya que a menudo incluye diferentes tipos de contenido sexual que muestra a menores, así como una amplia gama de servicios en línea con posibilidades muy diferentes y arquitecturas técnicas subyacentes. Aunque el tema ha recibido una amplia atención de los medios internacionales, la literatura académica sobre la transmisión en vivo y el abuso infantil sigue siendo muy poca y escasamente desarrollada: Un artículo de revisión de 2024 encontró solo 8 artículos revisados por pares sobre el tema ([Drejer, Riegler, et al., 2024](#)).

Estos artículos incluso suelen enfocarse en una modalidad específica de tráfico sexual transnacional facilitado o en línea, con un foco predominante en Filipinas ([Drejer, Riegler, et al., 2024](#)). El escenario general es que los niños en contextos de bajos ingresos se ven involucrados en varias formas de trabajo sexual y de cámaras web para menores de edad, posiblemente presionados por miembros de la familia o en busca de una forma de sobrevivir en un contexto de pobreza desesperada, con dinámicas de coerción poco claras ([Christensen y Woods, 2024](#)). Diversas investigaciones no revisada por pares publicada en colaboración con las fuerzas del orden ha descrito algunos de los aspectos de este fenómeno, donde las “transmisiones” en vivo mediadas a través de plataformas de videollamadas se venden a audiencias internacionales ([Teunissen et al., 2021](#); [Teunissen & Napier, 2023](#)). Una revisión de los 30 casos públicos procesados por este tipo de abuso sexual en línea en el Reino Unido entre 2013 y 2022 muestra que la coordinación de estas transacciones puede ocurrir a través de una multitud de canales, incluidos foros de citas en línea, sitios de contenido para adultos y aplicaciones de mensajería entre pares (peer-to-peer) ([Celiksoy et al., 2023](#)). Los materiales de prensa emitidos por Europol se han referido a este tipo de actividad como “abuso infantil en vivo y a distancia”, y este es el foco central del trabajo de varias organizaciones no gubernamentales sobre el tráfico sexual internacional y el abuso y la explotación sexual infantil en línea ([Europol, 2024](#)).



Una segunda modalidad menos desarrollada de abuso sexual infantil, relacionada con el intercambio de videos en tiempo real, tiene menos que ver con la coerción directa y más con la manipulación indirecta y el engaño. Este trabajo se caracteriza por centrarse en la noción emergente de “grooming” en línea (Salter & Sokolov, 2024). El concepto se utiliza cada vez más en el contexto “en vivo” para describir la actividad en la que los jóvenes activos en varios espacios en línea (por ejemplo, juegos en línea multijugador, foros en línea) son abordados por otros usuarios con el objetivo de desarrollar una relación con ellos y luego, a la larga, convencerlos o coaccionarlos para que se unan a videollamadas donde se les presiona para realizar una actividad sexual (Drejer, Sabet *et al.*, 2024). Este contenido se considera “autogenerado” porque generalmente lo producen los propios jóvenes, y gran parte de la cobertura de este tipo de “grooming” se centra en contextos de altos ingresos “cercanos” en lugar de los países internacionales de bajos ingresos que están en el centro de la conversación sobre el “abuso infantil en vivo y distante” (Vallance, 2024).

Sin embargo, la dinámica de la producción de contenido sexual autogenerado por menores puede ser ambivalente, especialmente cuando se trata de jóvenes mayores y sus importantes y legítimos esfuerzos por lograr su autonomía sexual en línea (Quayle, 2022). Para complicar aún más las cosas, múltiples factores sociales y económicos complejos motivan la creación de diversas formas de contenido sexual autogenerado por menores de edad (Cooper *et al.*, 2016). Como lo han destacado algunas investigaciones recientes realizadas por expertos en políticas y seguridad de Internet, las principales plataformas de contenido generado por el usuario, como Instagram, se han convertido recientemente en el hogar de usuarios menores de edad que venden o intercambian material sexual que los muestra a ellos mismos (Thiel *et al.*, 2023).

Actualmente no existen datos de incidencia multiplataforma sólidos e independientes sobre la prevalencia de este tipo de material sexual de menores “autogenerado” en el contexto de transmisión en vivo. Sin embargo, la cobertura sustancial de las transmisiones en vivo y los problemas de abuso sexual infantil en los medios internacionales en los últimos años ha resaltado algunas de las dinámicas potenciales en juego, incluso en una variedad de servicios de plataformas públicas populares que los jóvenes en los Estados Unidos y más allá usan con frecuencia. Una investigación de alto perfil realizada por Bloomberg, en 2022, indagó sobre la forma en que los jóvenes streamers de Twitch estaban siendo acosados, rastreados y, en algunos casos, obligados, convencidos o engañados para realizar actos sexuales (D’Anastasio, 2022). Se han vinculado informes de “sextorsión” a ciertos servidores de Discord (Boburg *et al.*, 2024; Goggin, 2023). La propia investigación interna de TikTok sugirió que los niños se desnudaban en su servicio TikTok Live a cambio de regalos en línea (Allyn *et al.*, 2024).

Periodistas han documentado cómo las posibilidades de ciertas plataformas han facilitado una mayor difusión de material de abuso sexual infantil. Los streamers jóvenes pueden creer que las transmisiones son completamente efímeras, pero algunas plataformas tienen funciones integradas que permiten guardar fragmentos de las transmisiones y hacerlos visibles después del hecho. Para una investigación de Bloomberg, de 2024, los periodistas trabajaron en alianza con el Centro Canadiense para la Protección de la Infancia para investigar el archivo de “Clips” de Twitch, videos cortos de hasta 20 segundos que habían sido producidos por espectadores de sus momentos de transmisión favoritos, y descubrieron que un número sorprendentemente grande (7,5% de una muestra de 1100 “clips”) podría ser clasificado como que exhiben representaciones sexualmente explícitas de menores (D’Anastasio, 2024; Winslow, 2024). Los informes sugieren que al menos algunos contextos de transmisión en vivo pueden implicar la reproducción de material sexual previamente obtenido que muestra a menores en foros de video y audio en vivo. Una investigación de la NBC sobre Discord sugirió un complejo ecosistema de actores maliciosos que incluía “cazadores” que localizaban a chicas y las invitaban a un servidor de Discord, ‘conversadores’ que eran responsables de chatear con las chicas y seducirlas, y los ‘loopers’ que transmitían contenido sexual previamente grabado y se hacían pasar por menores para alentar a las chicas reales a participar en actividades sexuales” (Goggin, 2023).

Algunos informes e investigaciones que analizan las transmisiones en vivo también han abordado formas de acoso y “troleo” (acción de “publicar mensajes provocativos, ofensivos o fuera de lugar con el fin de boicotear algo o a alguien, o entorpecer la conversación”, El Diccionario de la Lengua Española, 2024) en red en tiempo real, donde usuarios maliciosos inundan nuevos canales y buscan causar problemas a ciertos streamers (Han *et al.*, 2023), especialmente mujeres y personas de comunidades racializadas (Jackson, 2019; Ruberg, 2021). Este acoso puede potencialmente implicar el intercambio de material de abuso sexual infantil, ya sea para infligir daño a potenciales espectadores o para “destruir” ciertas transmisiones y lograr que los moderadores o la detección automática de contenido de la plataforma las cierren. Periodistas de 404 Media han demostrado cómo esta estrategia ha sido utilizada recientemente por varios grupos de hackeo y fraude para cerrar servidores rivales de Discord (Cox, 2024a).

Alcance y Métodos de Denuncia

En general, en este pequeño pero creciente cuerpo de investigación e investigación periodística, el término “transmisión en vivo / livestreaming” se utiliza para referirse a al menos siete categorías de plataformas relacionadas, pero distintas:

- Redes sociales importantes y de propósito general que tienen productos o “espacios” de transmisión en vivo destinados a transmisiones públicas o para una gran audiencia (por ejemplo, TikTok Live, Instagram Live, Facebook Live)
- Plataformas específicas para compartir videos en vivo que comenzaron principalmente en el contexto de juegos en línea, pero que ahora albergan una amplia gama de entretenimiento y comentarios (por ejemplo, Twitch, Kick, Discord)
- Plataformas de transmisión de audio o video que se basan en la idea de un evento en vivo transmitido internacionalmente, como una charla o un concierto (por ejemplo, Clubhouse, Spotify Live)
- Plataformas de videoconferencia/videollamada de uso general (por ejemplo, Zoom, Teams, Skype, Jitsi, Webex)
- Plataformas de mensajería directa entre pares que también tienen funcionalidad de llamadas en vivo directas o en grupos pequeños (por ejemplo, Facetime Video, WhatsApp, Signal, Telegram)
- Aplicaciones de “chat de video aleatorio” al estilo de ChatRoulette que emparejan a dos usuarios aleatorios para una videollamada, generalmente como aplicaciones de navegador web (por ejemplo, Shagle, ChatRandom, ChatHub y otros servicios que siguen el ejemplo de Omegle/Chatroulette)
- Sitios web o aplicaciones diseñados específicamente para la actividad sexual de adultos y pornografía “en tiempo real” (por ejemplo, StripChat, Chaturbate)

Las investigaciones existentes también sugieren varias modalidades relacionadas, aunque distintas, de producción y difusión de contenido sexual infantil en lo que respecta a estos diferentes servicios. Estas incluyen:

- “Abuso infantil en vivo y a distancia” o tráfico sexual facilitado por videos en tiempo real (donde la investigación a menudo se centra en espectadores de transmisiones en vivo en contextos de altos ingresos con facilitadores en jurisdicciones de bajos ingresos)
- Material de abuso sexual infantil no en vivo (videos o imágenes) “retransmitido” a través de servicios de transmisión en vivo o distribuido a través de espacios de transmisión en vivo complementarias (por ejemplo, a través de enlaces en un chat que acompaña a una transmisión en vivo)

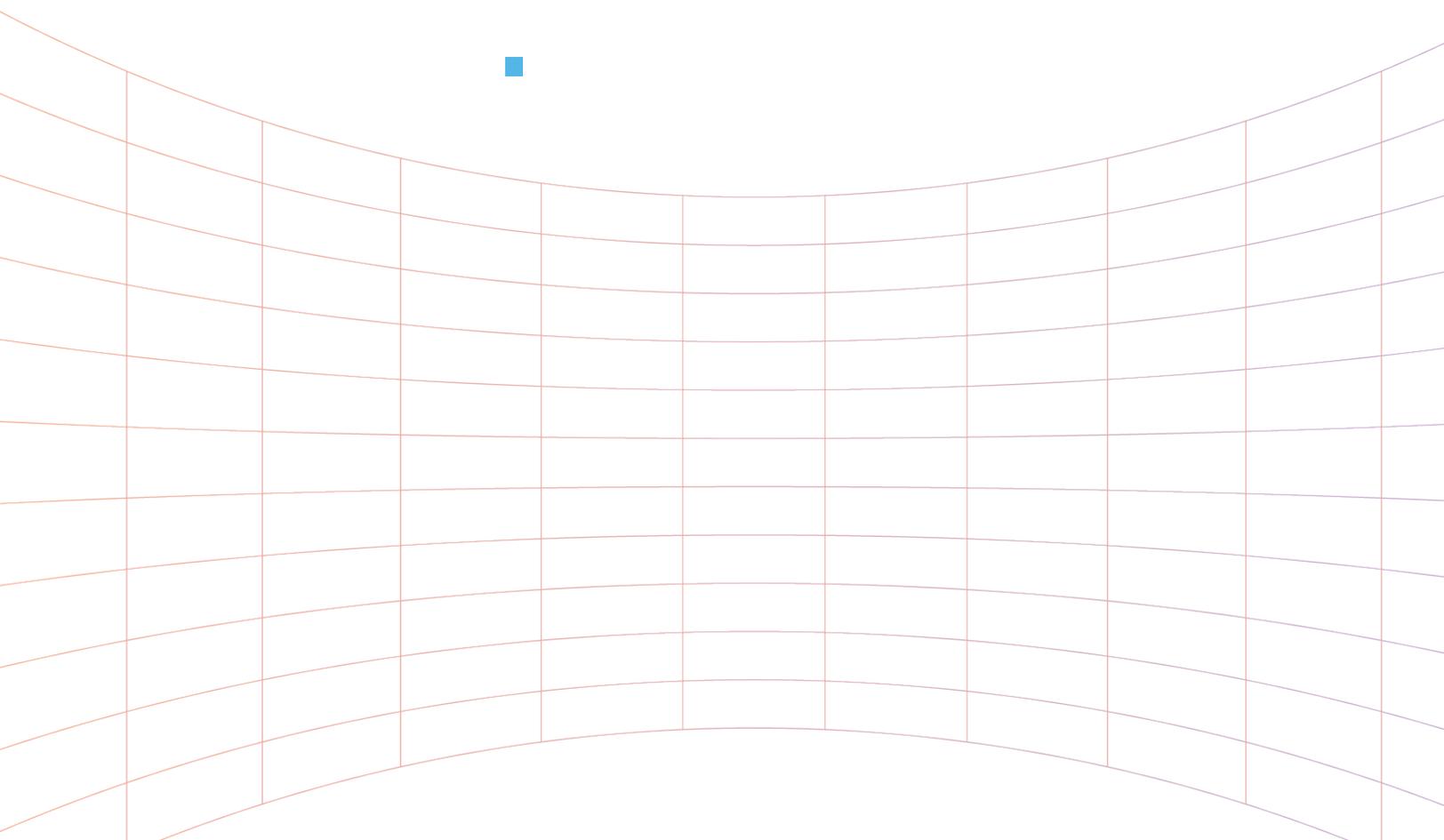
- Contenido sexual autogenerado por menores (donde la principal preocupación de los investigadores son los creadores en jurisdicciones de altos ingresos)
 - Creado y difundido a través de una transmisión en vivo (privada o pública)
 - Difusión en un entorno no presencial después de interacciones sociales en un servicio de transmisión en vivo (privado o público)

Los siete tipos diferentes de plataformas son suficientemente diferentes en su funcionalidad básica y arquitecturas técnicas como para que ningún proyecto de investigación pueda abordarlas todas exhaustivamente. Para complicar las cosas, los datos sobre la prevalencia real de las diferentes subcategorías de actividad siguen siendo escasos y de calidad limitada. Las organizaciones no gubernamentales han realizado algunos esfuerzos para medir la prevalencia de la explotación y el abuso sexual infantil, incluido el uso de transmisiones en vivo en lugares como Filipinas ([International Justice Mission y Laboratorio de Derechos de la Universidad de Nottingham, 2023](#)). Una operación policial europea coordinada en el verano de 2024 trabajó con 12 años de datos de “redes criminales que explotan sexualmente a niños en Filipinas”, analizando información sobre aproximadamente 12.000 cuentas que podrían estar vinculadas a 197 personas en el Espacio Económico Europeo, Reino Unido y Estados Unidos ([Europol, 2024](#)). Los informes de transparencia de los servicios populares de transmisión en vivo no distinguen entre contenido sexual autogenerado por menores y otras formas de material de abuso sexual infantil que se comparten en sus servicios. Por ejemplo, el informe de Twitch de 2023 afirma que la empresa tomó medidas en 12.801 casos en la primera mitad de 2023 en el marco de su “Política de Seguridad de los Menores de Edad”, que incluye “material ilegal de explotación y abuso sexual infantil, así como contenido y material que no es ilegal, pero que viola nuestras Directrices de la Comunidad al poner en peligro a menores” ([Twitch, 2023](#)). No está claro si estos casos involucran fragmentos de contenido, cuentas eliminadas o ambos.

Nuestro objetivo en este informe no es evaluar la prevalencia de la explotación y el abuso sexual infantil en la transmisión en vivo, sino examinar cómo las empresas y otros abordan el problema. Para ello, primero realizamos una revisión de la documentación disponible públicamente divulgada por las principales plataformas en las siete categorías analizadas anteriormente, en particular aquellas que tienen funcionalidad para productos de transmisión en vivo de “difusión/transmisión” pública. También revisamos las prácticas de sitios de pornografía, proveedores de llamadas privadas u otros servicios, e integramos información relacionada con sus esfuerzos de confianza y seguridad de estas empresas cuando fue posible.

Esto implicó el análisis de informes y documentos técnicos, comunicados de prensa, publicaciones de blogs y otros tipos de material de la industria publicados en línea por empresas, así como por asociaciones emergentes como Tech Coalition y proveedores externos de “tecnología de seguridad” que trabajan tanto con gobiernos como con diferentes plataformas. También realizamos 15 entrevistas con líderes de la industria, proveedores y expertos de la sociedad civil, pidiendo a estas personas que reflexionaran sobre sus herramientas y prácticas actuales (cuando corresponde), así como sobre los desafíos que enfrentan en sus esfuerzos diarios para salvaguardar las transmisiones. Los conocimientos adquiridos a partir de estos debates se perfeccionaron en un taller de medio día de duración con la participación de múltiples partes interesadas organizado por el CDT en junio de 2024.

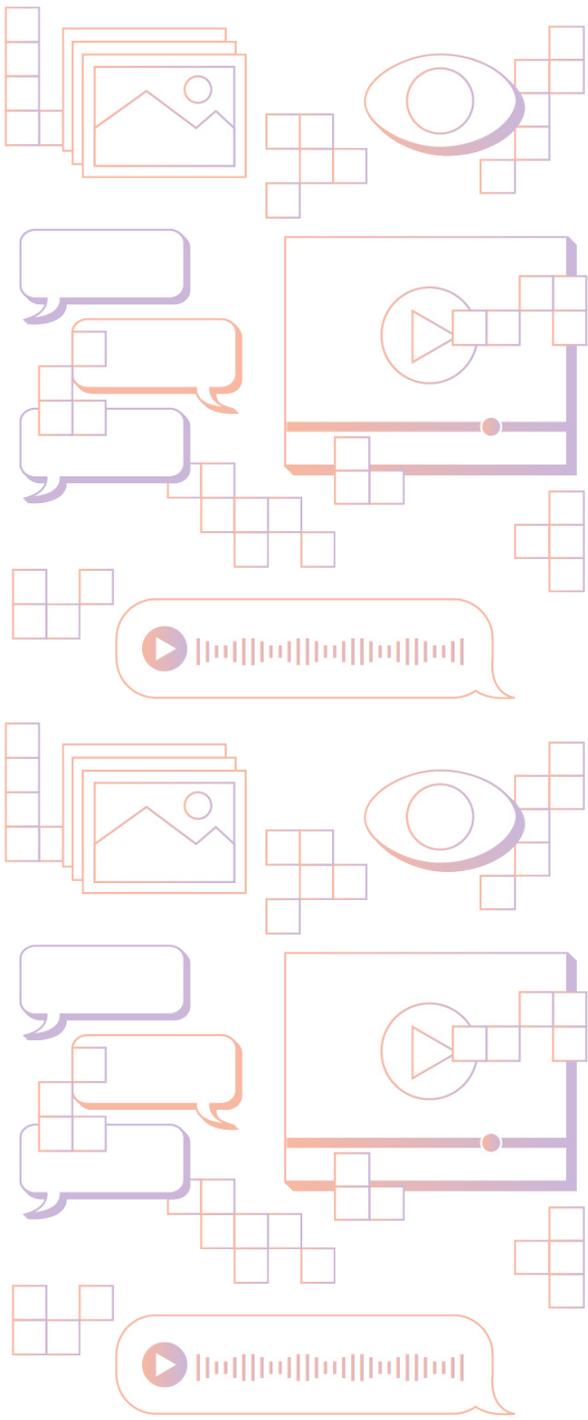
Como se señaló anteriormente, nuestro análisis se centra en la explotación y el abuso sexual infantil, que incluye material/imágenes de abuso sexual infantil. Este último está sujeto a un régimen jurídico y de gobernanza bien establecido. Sin embargo, la explotación y el abuso sexual infantil también incluye problemas como el grooming. Los investigadores, por ejemplo en estudios críticos sobre derecho y sexualidad, distinguen entre “conducta” y “contacto” (Baines 2019; McAlinden, 2006), que alude a una actividad como el grooming, que no es inherentemente violento o ilegal pero que podría conducir a futuros abusos o violencia bajo ciertas condiciones. Nuestra investigación mostró que las empresas están implementando cada vez más, al menos en cierto grado, medidas de confianza y seguridad que buscan contrarrestar este tipo de contacto potencialmente inseguro y, por lo tanto, también hemos incluido estas prácticas en el alcance de esta investigación cuando ha sido posible.



II. Protección de las Transmisiones: Una Visión General de la Confianza y la Seguridad

Las empresas que operan servicios en las siete categorías de “transmisión en vivo / livestreaming” antes mencionadas han implementado activamente diversas herramientas y prácticas para hacer cumplir sus políticas que prohíben el contenido sexual que muestre a menores. En general, las empresas con servicios de transmisión en vivo prohíben en sus plataformas la explotación y el abuso sexual infantil y material de abuso sexual infantil. Esto suele explicarse en los términos de servicio de la plataforma o en las directrices de la comunidad, con prohibiciones contra contenido sexual relacionado con jóvenes, incluido contenido que represente acciones o comportamientos específicos. Algunas plataformas (por ejemplo, Meta) ofrecen orientación específica contra el contenido que incluye la explotación y el abuso sexual infantil y material de abuso sexual infantil (Meta, 2024a). Los términos de servicio de las plataformas también suelen explicar que las empresas se reservan el derecho de tomar acciones contra dicho contenido o las cuentas que lo difunden. Estas actividades son parte de sus operaciones de confianza y seguridad (o T&S, por sus siglas en inglés), un término cada vez más utilizado por las empresas de tecnología para referirse a la elaboración de normas, aplicación de normas y diseño de sistemas que se implementan para prevenir o controlar tipos de comportamiento de los usuarios que van en contra de sus políticas o leyes locales (Caplan, 2023; Denyer Willis, 2023).

El trabajo académico sobre la gobernanza de contenido en ciertas plataformas de transmisión en vivo populares como Twitch ha identificado algunas de las ventajas clave que hacen que la implementación sea efectiva y la aplicación de políticas de contenido en entornos de video en tiempo real es potencialmente más difícil en comparación con otros productos de plataforma. Una característica importante del contexto en vivo es su cualidad de efímero, a menos que el streamer o alguien de la audiencia esté usando una herramienta (generalmente de terceros) para grabar la transmisión de alguna manera, la actividad que ocurre en el video de la transmisión puede desaparecer inmediatamente y ya no es accesible para la audiencia (Cai & Wohn, 2021), o, en general, ni para la plataforma misma, dependiendo de la arquitectura de la plataforma, las limitaciones de almacenamiento y otros factores. Esta característica no solo impide la posibilidad de futuras investigaciones por parte de las autoridades en casos graves (Horsman, 2018), pero también hace que sea más difícil para los empleados de confianza y seguridad y para los moderadores de la comunidad de una plataforma investigar y sancionar adecuadamente a los usuarios infractores según sea necesario.



En general, estas diversas intervenciones se pueden dividir en tres categorías amplias, que ocasionalmente pueden superponerse pero que, no obstante, proporcionan una estructura útil para reflexionar en las intervenciones de la plataforma: Métodos orientados al “diseño” que estructuran ciertas características del servicio para prevenir formas potenciales de abuso, métodos de “análisis de contenido”, que examinan el contenido real de una transmisión; y métodos de “señales”, que usan metadatos y patrones de comportamiento de cuentas para fundamentar decisiones de moderación en lugar de mirar el contenido real.

Una segunda posibilidad relacionada es rapidez—Cuando se produce una actividad violatoria en tiempo real, la respuesta de una plataforma debe ser extremadamente rápida para evitar la difusión del contenido en cuestión. Incluso en un contexto de transmisión en vivo más benigno (por ejemplo, una transmisión en la que un jugador popular está jugando ajedrez en línea con miles de espectadores), la interacción en vivo y una audiencia activa implican que las intervenciones de gobernanza a posteriori son menos efectivas y, además, potencialmente más difíciles de implementar (Cai & Wohn, 2019). Un conjunto activo de investigaciones sobre la interacción entre humanos y computadoras ha documentado las diversas herramientas y estrategias que los productores de contenido utilizan para moderar sus transmisiones, incluso en condiciones difíciles, como “brigadas” de usuarios que buscan acosar a ciertos streamers, mientras el streamer y los moderadores voluntarios afiliados intentan bloquear esas cuentas y eliminar sus comentarios (ver Brewer *et al.*, 2020; Cai *et al.*, 2023; Xiao, 2024). En un contexto en el que existe una amenaza a la seguridad pública o personal, como actos de violencia en vivo o autolesiones, las respuestas de confianza y seguridad no solo deben ser rápidas, sino también capaces de escalar hacia acciones fuera de la plataforma que involucre a las fuerzas del orden, funcionarios de salud pública u otros actores (Peralta, 2023; Zornetta & Pohland, 2022).

Por último, los medios de transmisiones en vivo son, al menos hasta cierto punto, técnicamente inescrutables para operaciones de confianza y seguridad de la plataforma. La mayoría de las transmisiones en vivo, por definición, contienen video y audio y, por lo tanto, son inherentemente más difíciles de analizar que las publicaciones clásicas basadas en texto o imágenes. Aunque la industria lleva tiempo desarrollando medios automatizados para analizar una variedad de multimedia, incluido el video, en busca de posibles violaciones de políticas a gran escala (Cobbe, 2021; Gorwa *et al.*, 2020), estas técnicas se desarrollaron para implementarse “en reposo” en instancias de contenido almacenadas, en lugar de implementarse en material en tiempo real y en constante cambio (Shenkman *et al.*, 2021). El análisis de contenido es inevitablemente mucho más complicado para los servicios cifrados de extremo a extremo, como las videollamadas o las videoconferencias privadas entre pares (Kamara *et al.*, 2021). En general, esto hace que la gestión de la transmisión en vivo sea un problema “resistente a la rutina” para las empresas de plataformas (Gorwa & Veale, de próxima publicación): uno en el que las posibilidades de espacios en vivo posiblemente requiera el desarrollo de técnicas especializadas que vayan más allá de sus operaciones habituales sobre material “frío” o “in situ”.

Si bien no existe un análisis sistemático publicado de las prácticas de confianza y seguridad en el contexto de la transmisión en vivo, nuestro estudio de este panorama revela una gran cantidad de herramientas y prácticas relacionadas que se han implementado o están en desarrollo activo. Algunas de ellas son estrategias y sistemas nuevos y especializados que han sido desarrollados por algunas empresas o proveedores para abordar específicamente la explotación y el abuso sexual infantil. Otras son simplemente nuevas implementaciones de técnicas clásicas generalmente utilizadas para detectar contenido sexual o desnudez, o para prevenir spam, fraude y otras formas de abuso. En general, estas diversas intervenciones se pueden dividir en tres categorías amplias, que ocasionalmente pueden superponerse pero que, no obstante, proporcionan una estructura útil para reflexionar en las intervenciones de la plataforma: Métodos orientados al “diseño” que estructuran ciertas características del servicio para prevenir formas potenciales de abuso, métodos de “análisis de contenido”, que examinan el contenido real de una transmisión; y métodos de “señales”, que usan metadatos y patrones de comportamiento de cuentas para fundamentar decisiones de moderación en lugar de mirar el contenido real. Un resumen está disponible en la Tabla 1.

Herramientas y Prácticas de Confianza y seguridad	Detalles
Diseño	Establecer barreras que obliguen a los usuarios a pasar por obstáculos adicionales antes de poder transmitir en vivo y, de ese modo, buscar reducir la probabilidad de que los streamers difundan material ilegal.
Umbral de la Cuenta	Requisitos para cumplir con ciertos criterios (por ejemplo, la cuenta debe tener una cierta cantidad de días de antigüedad o debe tener una cierta cantidad de suscriptores) para comenzar a transmitir en vivo
Verificación de Edad	Exige a los usuarios cuyas cuentas transmitan en vivo que verifiquen su edad de alguna manera; esto puede ser muy ligero (por ejemplo, autodeclaración de edad) o muy estricto (carga obligatoria de documento de identificación emitido por el gobierno)
Verificación Adicional	Exige a los usuarios cuyas cuentas transmiten en vivo que verifiquen algún aspecto de su identidad solicitando más que solo una cuenta de correo electrónico, como un número de teléfono o una tarjeta de crédito
Análisis de Contenido	Sistemas de detección de contenido automatizados que analizan el contenido (por ejemplo, video, audio o texto) de una transmisión en vivo.
Coincidencia: No en espacios en vivo	Un sistema de detección que intenta aplicar la tecnología de hash CSA estándar de la industria a contenido no en vivo (por ejemplo, fotos de perfil, fondos y otro contenido cargado por el usuario)
Coincidencia: Muestra basada en Hash	Un sistema que toma muestra de fotogramas de transmisiones en vivo y luego los ejecuta a través de herramientas de comparación de material de abuso sexual infantil para intentar detectar material "conocido"
Predictivo: Clasificador general de Seguridad (Video)	Clasificadores que buscan detectar la probabilidad de que una transmisión (o un fotograma dentro de una transmisión) contenga contenido que infrinja una regla, no limitado a la explotación y el abuso sexual infantil (por ejemplo, clasificadores de desnudez, estimación de edad)
Predictivo: Clasificadores CSA específicos (Video)	Clasificadores que buscan detectar la probabilidad de que una transmisión (o un fotograma dentro de una transmisión) contenga contenido de explotación y abuso sexual infantil en particular, tal vez a través de una versión optimizada de un clasificador de contenido sexual adyacente, una combinación de clasificadores (por ejemplo, detección de contenido sexual y categorización predictiva de la edad) o un sistema especialmente desarrollado
Predictivo: Clasificación de Audio	Busca detectar patrones de abuso en formas de onda de audio
Predictivo: Análisis de Texto	Transcripción de audio de transmisiones en vivo y su uso para ejecutar clasificadores de riesgo predictivos
Análisis de Texto de Chat	Análisis predictivo de comentarios publicados en una transmisión en vivo; palabras clave y otros indicadores potenciales para alertar

▲ **Tabla 1:** Resumen de ejemplos comunes de herramientas y prácticas de la industria para abordar la explotación y el abuso sexual infantil en transmisiones en vivo.

Herramientas y Prácticas de Confianza y seguridad	Detalles
Señales	Formas de análisis manuales o automatizados que utilizan datos no relacionados con el contenido para rastrear cuentas sospechosas o formas de comportamiento.
Indicadores de comportamiento de la Cuenta	Intercambio de señales y metadatos sobre “actores de riesgo” confirmados o pronosticados
Respuesta Jerarquización	Uso de señales o indicadores de usuario para responder a posibles violaciones de políticas en transmisiones

- ▲ **Tabla 1 (continuación)** : Resumen de ejemplos comunes de herramientas y prácticas de la industria para abordar la explotación y el abuso sexual infantil en transmisiones en vivo.

Diseño: Herramientas de Verificación, Autenticación y Moderación de la Comunidad

Los proveedores de servicios digitales de todo tipo han implementado desde hace mucho tiempo diversas medidas de propósito general para contrarrestar el fraude, el correo no deseado y otras formas de abuso potencial (Brunton, 2013). Técnicas como la autenticación de dos pasos (2FA, que requiere que los usuarios potenciales vinculen su cuenta a un número de teléfono activo, por ejemplo) son ampliamente utilizadas por las redes sociales no solo para dificultar la creación de cuentas fraudulentas, sino también como una posible protección contra ataques de phishing e intrusión de cuentas (Tirfe & Anand, 2022). Es al menos conceptualmente posible, aunque no está demostrado empíricamente, que quienes buscan crear o difundir diversas formas de contenido sexual de menores de edad, sean menos propensos a hacerlo si su cuenta está vinculada a un número de teléfono u otra forma de verificación complementaria, aunque los actores maliciosos comprometidos pueden eludir los requisitos de 2FA a través de números de teléfono desechables, intercambio de SIM y otras técnicas.

De manera similar, varias características de diseño aparentemente mundanas (muchas de las cuales se subcontratan a los streamers y sus moderadores designados, quienes pueden optar por implementar varias formas de filtrado de chat, “moderación automática” o fricción del espectador) pueden desempeñar un papel en el combate estructural de múltiples formas de abuso. Por ejemplo, los streamers de Twitch pueden

establecer “requisitos de verificación a nivel de canal”, bloqueando a los espectadores que no hayan verificado sus correos electrónicos o números de teléfono, o que no hayan cumplido otras condiciones ([Twitch, sin fecha](#)). En ciertas plataformas, los streamers pueden bloquear enlaces compartidos y acceder a listas de bloqueo de ciertas palabras o detección y bloqueo de “toxicidad de chat” en tiempo real de varias formas, algunas de las cuales son administradas por la comunidad, mientras que otras son herramientas proporcionadas directamente por las plataformas para su fácil integración e implementación.

Otro tipo de umbral implementado por algunas empresas en sus plataformas de transmisión en vivo involucra medidas relacionadas con la popularidad de una cuenta. YouTube, por ejemplo, había implementado previamente una política según la cual todas las cuentas que quisieran hacer transmisiones en vivo debían tener al menos 50 suscriptores en el canal, una forma simple de fricción destinada a evitar que un actor malicioso cree impulsivamente un canal de YouTube y transmita en vivo un acto de violencia, por ejemplo. Según nuestras entrevistas en la industria, estos umbrales generales de cuenta se implementaron principalmente para evitar la transmisión en vivo de autolesiones y otros actos de violencia (por ejemplo, el tiroteo de Christchurch), pero también para dificultar considerablemente la transmisión de la mayoría de las formas de contenido sexual que muestra a menores.

El umbral de YouTube fue ampliamente difundido en blogs de marketing de influencers ([Tommy I., 2023](#)), pero no está claro si esto se aplica sólo a las transmisiones en vivo desde dispositivos móviles ([Google, 2024a](#)). Sin embargo, existen medidas de seguridad adicionales en torno a las transmisiones en vivo: la última documentación de YouTube establece que para realizar una transmisión en vivo, uno debe autenticar un número de teléfono y, para “funciones avanzadas”, como incrustar transmisiones en vivo, uno debe “construir un historial de canal suficiente” o completar un proceso de identificación de documento personal o identificación biométrica ([Google, 2024b](#)). En TikTok, aunque no parece haber un umbral similar para salir en vivo ([TikTok, 2024e](#)), hay umbrales para acceder a herramientas específicas como Live Studio (por ejemplo, los usuarios de Estados Unidos deben tener al menos 10.000 seguidores) ([TikTok, 2024c](#)).

MEDIDAS ESPECÍFICAS DE SEGURIDAD PARA NIÑOS

Algunas plataformas de transmisión en vivo también están creando cada vez más umbrales de edad específicos destinados a evitar que los jóvenes violen sus políticas de diversas maneras. Las directrices de la comunidad de TikTok señalan explícitamente que los usuarios menores de 18 años no pueden realizar transmisiones utilizando la función TikTok LIVE de la plataforma, mientras que para Twitch, los usuarios deben tener 13 años o más ([TikTok, 2024e](#); [Twitch, 2024](#)). Debido a la ley de Estados Unidos de Protección de la Privacidad en Línea para Niños, de 1998, (COPPA, por sus siglas en inglés) los servicios en línea enfrentan reglas especiales sobre cómo manejan los datos de personas menores de 13 años ([Reyes *et al.*, 2018](#)). La ley ayudó a dar origen

al formulario de registro de cuentas de Internet que consiste en “autoinformar tu edad”, pero cada vez más plataformas parecen estar implementando también medidas adicionales para predecir las edades de algunos de sus usuarios. Twitch, por ejemplo, utiliza indicadores no revelados para entrenar modelos que intentan “capturar y cancelar cuentas pertenecientes a usuarios menores de 13 años, así como bloquear a usuarios previamente suspendidos por ser menores de 13 años para que no creen nuevas cuentas” (Twitch, 2022). La compañía también ha implementado algunas medidas donde las cuentas clasificadas como “potencialmente vulnerables” a través de este tipo de métricas predictivas deben pasar por “requisitos obligatorios de verificación telefónica antes de poder transmitir en vivo”.

TikTok afirma de manera similar que utiliza un conjunto de “enfoques adicionales no revelados para identificar y eliminar a los titulares de cuentas que se sospecha que son menores de edad”, analizando los patrones de comportamiento de las cuentas para hacer predicciones sobre los usuarios menores de 18 años. Los usuarios marcados por estos modelos predictivos deben someterse a un proceso obligatorio de verificación de edad si desean transmitir en vivo, lo que implica compartir “una selfie [...con] una identificación emitida por el gobierno y (b) un trozo de papel que indique de forma clara y legible un código único” (TikTok, 2024a). El informe de transparencia de TikTok afirma que en los primeros dos trimestres de 2024 la compañía eliminó un promedio de 21 millones de cuentas como sospechosas de ser menores de 13 años a través de técnicas automatizadas, aunque no revela cuántas cuentas fueron marcadas por este tipo de verificación secundaria de edad en el contexto específico de intentar transmitir en vivo (TikTok, 2024d). Nuestras entrevistas con la industria sugirieron que las empresas utilizan el gráfico social de los usuarios, los tipos de temas y transmisiones que ven y otras formas no reveladas de metadatos para construir estos modelos de clasificación por edad, pero hay poca información pública concreta sobre cuán exhaustivos y efectivos son realmente estos esfuerzos.

La creciente tendencia hacia determinadas formas de verificación de la edad en el contexto de la transmisión en vivo pública tiene algunos puntos en común con los requisitos legales en la industria de contenido para adultos. En la Sección 2257 del Título 18 del Código de los Estados Unidos se exige que los sitios de pornografía con sede en Estados Unidos mantengan registros que confirmen que todos sus artistas son mayores de 18 años. Este tipo de requisitos se han implementado ampliamente por muchos organizadores internacionales de contenido para adultos. Por ejemplo, el popular sitio de cámaras con sede en Chipre, Stripchat, designado por la Comisión Europea en diciembre de 2023 como una Plataforma en línea muy grande por los requisitos especiales de la Ley de Servicios Digitales, tiene una regla a partir del verano de 2024, que establece que todos los nuevos artistas deben presentar pasaportes, documento nacional de identidad u otra identificación para demostrar que son mayores de 18 años antes de que se les permita crear contenido (Stripchat, 2024).

Los sistemas de verificación de identidad también plantean importantes riesgos de privacidad y seguridad: Desde hace tiempo existen preocupaciones sobre la posibilidad de que los proveedores de verificación de terceros se conviertan en puntos vulnerables de falla y/o en objetivos de los piratas informáticos a medida que se apoderan cada vez más de datos personales confidenciales

En nuestra entrevista con un sitio de gran dimensión para adultos, el personal destacó cómo implementan medidas de verificación de terceros (utilizando herramientas de un gran proveedor externo) de todos los artistas, incluso en sitios subsidiarios que tienen funcionalidad de transmisión en vivo. Para ayudar a garantizar que los streamers mostrados fueran realmente los que habían pasado por el proceso de verificación de edad y del documento de identidad emitido por el gobierno, esta empresa afirmó tener moderadores humanos “siempre involucrados”, observando activamente cada transmisión para detectar violaciones de políticas y verificaban las identidades de los artistas contra los documentos registrados.

En el contexto de la transmisión pública en vivo, estas formas de verificación de identidad pueden dificultar la proliferación de algunas -pero no todas- las formas de material de abuso sexual infantil. Las cuentas pueden ser hackeadas y tomadas por actores maliciosos, los documentos pueden ser falsificados y los sistemas automatizados para la verificación de edad utilizados por las plataformas y sus contratistas externos de “garantía de edad” pueden cometer errores y plantear importantes preocupaciones sobre la privacidad. Los sistemas de verificación de identidad también plantean importantes riesgos de privacidad y seguridad: Desde hace tiempo existen preocupaciones sobre la posibilidad de que los proveedores de verificación de terceros se conviertan en puntos vulnerables de falla y/o en objetivos de los piratas informáticos a medida que se apoderan cada vez más de datos personales confidenciales (Blake, 2019; Persson, 2024). Un informe de 404 Media del verano de 2024 reportó que uno de los socios de verificación de TikTok, una empresa con sede en Israel llamada AU10TIX protegió incorrectamente sus bases de datos, lo que permitió a los investigadores de ciberseguridad acceder a las licencias de conducir y otra información personal de aquellas personas a las que se les exigió para utilizar los servicios de la empresa (Cox, 2024b).

Análisis de Contenido: Escrutinio de las Transmisiones en busca de Violaciones de Políticas

Las transmisiones en vivo son inherentemente difíciles de analizar para detectar violaciones de las políticas de contenido con herramientas automatizadas que funcionan con precisión y eficacia a escala. Sin embargo, eso no ha impedido que la industria pruebe diversas estrategias. Estas van desde enfoques clásicos para encontrar material de abuso sexual infantil “conocido” previamente confirmado, adaptado al contexto de video en tiempo real, hasta métodos predictivos más complejos que buscan descubrir con precisión material “nuevo” y previamente desconocido con modelos de visión por computadora. La industria también parece estar implementando de manera creativa varias herramientas de comparación y predicción durante las transmisiones sobre material complementario que no es parte directa del feed del streamer.

MATERIAL “CONOCIDO”: VIDEO Y AUDIO

Se ha revelado poco públicamente sobre las técnicas específicas que utilizan las plataformas para trabajar con video en vivo cuando se trata de la detección de posible contenido de material de abuso sexual infantil. Empresas como TikTok mencionan que pueden utilizar una amplia gama de tecnologías, “incluidos [sus] propios sistemas y software de comparación de hashes como PhotoDNA de Microsoft, Content Safety API de Google y CSAI Match de YouTube”, sin profundizar en las implementaciones específicas y las ventajas relativas de estos diversos sistemas (TikTok, 2024b). En entrevistas, algunos participantes de la industria mencionaron que se pueden usar técnicas como “hashing de video sensible a la escena” (SSVH) para colapsar videos en sus “fotogramas clave”, que luego se pueden ejecutar a través de sistemas de comparación de hash convencionales como Photo DNA.

Un pequeño número de empresas también parece estar implementando técnicas computacionalmente más livianas, como el hashing de audio, que se han vuelto relativamente robustas en los últimos años y han ganado popularidad a través de herramientas de comparación de “cuál es esta canción”, como Shazam. Aquí, la industria puede hacer uso de hashes de audio de material de abuso sexual infantil confirmado, para, al igual que con los enfoques SSHV, evitar que los streamers aprovechen el ofrecimiento de los en vivo para “reproducir” públicamente material de abuso sexual infantil previamente confirmado. Aunque en general se considera que estas tecnologías son rápidas y precisas, la preocupación de algunos entrevistados fue que no valía la pena el costo computacional para muchas empresas dada la percepción de rareza de este tipo de material de abuso sexual infantil “previamente conocido” transmitido públicamente por los streamers.

Según nuestras entrevistas, las empresas parecen estar usando activamente tecnologías de hash cuando el contenido es estático, en reposo, y puede escanearse y compararse para detectar posibles infracciones. Por ejemplo, las plataformas de streaming a gran escala que permiten a los usuarios cargar fondos, imágenes al chat, emojis personalizados u otros tipos de medios, bloquean sistemáticamente esas cargas para evitar la fácil difusión de material de abuso sexual infantil conocido. El hash de contenido no cifrado y contenido no en vivo es ampliamente implementado por los principales servicios de plataforma con distintos modelos de negocios (mensajería peer to peer, nube, carga de contenido generado por el usuario) como medida de protección general y mejor práctica.

MATERIAL “NUEVO”: VIDEO, AUDIO, TEXTO

También se están realizando esfuerzos más activos para analizar el contenido en el contexto único de transmisión en vivo utilizando nuevas técnicas que podrían ayudar a las plataformas a identificar material “previamente desconocido”: contenido creado en vivo en el lugar que, por definición, no se ha visto antes y, por lo tanto, no estaría contenido en bases de datos hash de material de abuso sexual infantil conocido. El punto de partida natural para este tipo de análisis es intentar trabajar con la transmisión de video en sí, utilizando varios modelos de visión computarizados para predecir la posibilidad de que se produzcan ciertas violaciones de políticas. Las grandes empresas utilizan cada vez más estos clasificadores en áreas adyacentes, como la detección de contenido sexual general. Por ejemplo, en diciembre de 2023, un comunicado de prensa de Instagram señaló que la empresa había implementado un “nuevo esfuerzo de cumplimiento automatizado” que aumentó cinco veces sus “eliminaciones automáticas de los Instagram Lives que contenían desnudez adulta y actividad sexual” ([Meta, 2023](#)).

Algunos proveedores externos ofrecen productos que buscan permitir que plataformas de diversos tamaños integren este tipo de soluciones de análisis de video en sus entornos de transmisión en vivo. Una herramienta comercializada por una gran empresa de soluciones de Confianza y Seguridad, por ejemplo, realiza un seguimiento consecutivo de fotogramas de un video en vivo y luego ejecuta la clasificación de imágenes en ellos. Luego, varios modelos predictivos devuelven puntajes para diversas “áreas de daño” potenciales, lo que permite a los clientes tomar acciones de moderación según puntajes específicos o una combinación de puntajes.

Algunos proveedores externos ofrecen productos que buscan permitir que plataformas de diversos tamaños integren este tipo de soluciones de análisis de video en sus entornos de transmisión en vivo. Una herramienta comercializada por una gran empresa de soluciones de Confianza y Seguridad, por ejemplo, realiza un seguimiento consecutivo de fotogramas de un video en vivo y luego ejecuta la clasificación de imágenes en ellos. Luego, varios modelos predictivos devuelven puntajes para diversas “áreas de daño” potenciales, lo que permite a los clientes tomar acciones de moderación según puntajes específicos o una combinación de puntajes. Un producto popular con documentación pública y características similares es Rekognition de Amazon, que ofrece detección de desnudez en imágenes y videos, y también permite a los clientes ajustarlo para fines personalizados ([Amazon Web Services, 2020](#)). Para poder detectar potencialmente contenido sexual que incluya a menores, una estrategia analizada por los proveedores implica combinar herramientas de estimación de edad con varios clasificadores de desnudez o conducta sexual (de adultos) disponibles públicamente. Los modelos resultantes se entrenan utilizando métodos de aprendizaje supervisados o no supervisados para intentar detectar contenido nuevo basándose en patrones de esos datos de entrenamiento.

Otra estrategia que está siendo implementada por una nueva generación de proveedores centrados en la seguridad infantil implica la creación de clasificadores específicamente para detectar contenido sexual de menores, modelos de entrenamiento en conjuntos de datos de material de abuso infantil confirmado que son proporcionados por las autoridades u organizaciones como la NCMEC o la Internet Watch Foundation. La idea aquí es utilizar conjuntos de datos de material de abuso sexual infantil para predecir la probabilidad de que una nueva imagen (o fotograma de video) también pueda ser material de abuso sexual infantil. Las herramientas publicitadas abiertamente con este tipo de funcionalidad incluyen ‘Safer Predict’ (Thorn, 2022) y el producto ‘HarmBlock’ (MSAB, 2023), este último afirma ser capaz de trabajar en un contexto de transmisión en vivo (Payt, 2024). Los activistas alegan que este tipo de enfoque es más prometedor que agrupar modelos problemáticos de estimación de edad con modelos de desnudez, aunque hay preguntas abiertas sobre cuán globalmente representativos son los conjuntos de datos de material de abuso sexual infantil de estos modelos, las cuestiones éticas en torno a poner estos conjuntos de datos a disposición de empresas privadas y su eficacia real a escala de plataforma dada la falta de evaluación comparativa pública y esfuerzos independientes para verificar estos modelos en varios entornos.

Otra técnica emergente implica alejarse del material de video real y, en cambio, trabajar con el audio utilizando modelos predictivos. Algunos entrevistados mencionaron sus esfuerzos para entrenar modelos predictivos en audio, argumentando que los videos de abuso sexual infantil exhiben formas de onda y patrones distintos que pueden usarse para luego encontrar videos similares. Sin embargo, no está claro cuán robustas y distintas son estas formas de onda y patrones, especialmente cuando se busca detectar comportamiento infractor en plataformas que también albergan amplias representaciones de violencia y sexualidad en, por ejemplo, videojuegos populares.

El enfoque más común parece funcionar con el audio transcrito de las transmisiones y el texto asociado con las transmisiones (por ejemplo, en el chat y otros metadatos de las transmisiones). Muchas plataformas grandes ya han integrado modelos de subtítulos de voz a texto en sus ofertas en vivo (lo que permite subtítulos en vivo multilingües, por ejemplo), lo que hace que sea relativamente sencillo para ellos usar estas transcripciones para otros tipos de acciones de moderación. Los empleados de las plataformas de transmisión en vivo públicas analizaron con nosotros técnicas de análisis de texto de diversa complejidad (que van desde simples “listas de alertas” basadas en palabras clave hasta grandes modelos de lenguaje ajustados) como una herramienta cada vez más utilizada para marcar transmisiones sospechosas para revisión humana. Sin embargo, no todo el mundo está haciendo esto. Algunos participantes mencionaron que deseaban que su empresa invirtiera en transcripción a gran escala, ya que les abriría muchas nuevas posibilidades de confianza y seguridad, pero que hacerlo se consideraba prohibitivo, de un precio muy alto, desde una perspectiva de computación y almacenamiento.

Por supuesto, el uso de modelos de aprendizaje automático para el análisis textual conlleva riesgos, incluidos la posibilidad de perder el contexto en el que se realiza una publicación, particularmente en entornos multiculturales, y el agravamiento de la discriminación contra grupos ya marginados (Duarte *et al.*, 2017). Los avances más recientes en modelos de lenguaje multilingües de gran escala que se pueden utilizar para evaluar contenido en diferentes idiomas aún plantean problemas para el análisis en idiomas de “bajos recursos” o aquellos para los que los datos de entrenamiento son escasos (Nicholas & Bhatia, 2023).

Trabajar con transcripciones o señales de forma de onda de audio puede complementar el modelo clásico de “marcado/señalización” de los informes de violaciones de las normas de la comunidad por parte de los usuarios comunes, un elemento básico de cómo los intermediarios buscan identificar y responder a las quejas (Crawford & Gillespie, 2016). Prácticamente todas las plataformas declaran públicamente que intentan responder a las alertas en las transmisiones en vivo lo más rápidamente posible, aunque cada vez más han comenzado a utilizar herramientas predictivas adicionales y formas de análisis de metadatos para ayudar a priorizar esta revisión y así poder responder a las denuncias más “urgentes” y de alto riesgo.

Una forma de hacer esto es creando un “puntaje de riesgo” de los tipos de interacciones en un chat o analizando otras formas de información proporcionadas por un streamer que no son necesariamente parte de la transmisión principal (por ejemplo, miniaturas de videos, títulos y otras palabras clave), o usando análisis basados en audio o video como indicadores para la observación humana. Thorn publicita un clasificador de texto predictivo que realiza este tipo de puntuación de riesgo, y sus materiales de marketing públicos incluyen una cita de un proveedor de servicios que afirma que el “clasificador de texto” de la empresa “mejora significativamente nuestra capacidad para priorizar y escalar cuentas y contenidos de alto riesgo. Las múltiples etiquetas y puntuaciones de riesgo ayudan a nuestro equipo a centrarse en las cuentas problemáticas, algunas de las cuales en las que ya desconfiábamos pero carecíamos de evidencias legales antes de implementar el clasificador” (Thorn, sin fecha). En teoría, las empresas pueden utilizar este tipo de clasificación de texto en transmisiones en vivo transcritas para marcarlas para revisión humana.

Algunas empresas también están utilizando métodos de análisis de texto para intentar predecir la incidencia de “grooming” y conductas relacionadas con el material de abuso sexual infantil, en lugar de solo el material de abuso sexual infantil real. Por ejemplo, el producto Safer by Thorn afirma que puede buscar “comportamiento de explotación sexual infantil” en texto (Thorn, 2022). Aunque las empresas se muestran públicamente vagas respecto de los detalles, parece como si múltiples plataformas estuvieran desarrollando listas de palabras clave que pueden usarse como indicadores de problemas de seguridad infantil en sentido amplio, a pesar del alto grado potencial de falsos positivos y los problemas potenciales en diferentes contextos lingüísticos. La Tech Coalition, junto con Thorn, ha creado un “Centro de Palabras Clave CSAM” que permite a las plataformas acceder y gestionar de forma comunitaria un conjunto de palabras clave en varios idiomas que se pueden utilizar para diversas intervenciones de confianza y seguridad (una breve sección de preguntas frecuentes públicas establece que, idealmente, el centro no debería utilizarse “para bloquear estrictamente palabras clave específicas que coinciden con la lista”, ya que “la fuerte preferencia es usar la lista como punto de partida para el entrenamiento de modelos de aprendizaje automático” que pueden marcar material para revisión o bloquear elementos de un chat) (Thorn, 2024).

Como lo expresó un representante de la industria entrevistado, los sistemas orientados a palabras clave pueden ser frágiles y, sin embargo, útiles. Si bien su equipo de seguridad con frecuencia se ve inundado de alertas que pertenecen a transmisiones que claramente no son problemáticas cuando, por ejemplo, una discusión activa entre fanáticos de una película de DC Comics, “El Escuadrón Suicida”, conduce a la proliferación repentina de transmisiones con palabras clave de autolesión en el título y la descripción, los sistemas de análisis de texto aún se pueden usar para analizar chats públicos, descripciones de transmisiones y títulos de transmisiones para guiar la revisión humana posterior.

Señales: Investigación, Seguimiento y Predicción de Comportamientos Violatorios

Un experto de la industria señaló que los esfuerzos para prevenir el material de abuso sexual infantil en las transmisiones en vivo se han vuelto más sofisticados en los últimos años y han desplazado el foco de la detección del contenido potencialmente infractor hacia la comprensión del comportamiento de los actores maliciosos confirmados o sospechosos. Desde esta perspectiva, una primera ola de esfuerzos en materia de confianza y seguridad buscó analizar videos en tiempo real, una segunda ola se inclinó hacia el uso de audio (y especialmente la transcripción), y la última, las prácticas

más vanguardistas de la industria implican una tercera ola de intervenciones basadas principalmente en “señales” del comportamiento de los actores. El último conjunto de prácticas es el resultado de un esfuerzo de varios años para tratar de desarrollar medidas no orientadas al contenido para la detección de material de abuso sexual infantil (por ejemplo, observando los metadatos de los archivos) (Pereira *et al.*, 2023), y aunque inicialmente no se desarrollaron para el contexto de transmisiones en vivo, estas herramientas también pueden ser una parte útil de las operaciones de confianza y seguridad de las plataformas de streaming.

Una forma en que se pueden utilizar las señales es para compartir datos sobre cuentas y actividades que violan las políticas de una plataforma con respecto a la explotación y el abuso sexual infantil. Esto puede socavar la capacidad de actores maliciosos conocidos para operar en múltiples plataformas. Un segundo enfoque es utilizar señales para predecir si ciertas actividades o cuentas están potencialmente involucradas en la distribución de explotación y abuso sexual infantil. La mayoría de las principales plataformas de transmisión en vivo emplearán ambos enfoques.

Un nuevo proyecto importante facilitado por el consorcio Tech Coalition que apoya enfoques basados en señales para prevenir la distribución de la explotación y el abuso sexual infantil es Lantern. Las plataformas participantes pueden compartir “señales”, como metadatos vinculados a cuentas que se haya confirmado que son distribuidores activos de material de abuso sexual infantil, para que otras plataformas los utilicen. Estos indicadores (los ejemplos proporcionados por Lantern incluyen direcciones de correo electrónico, nombres de usuario y ciertas palabras clave) se pueden utilizar para informar investigaciones realizadas por equipos especializados en las plataformas, o para alimentar modelos que puedan usarse para marcar (o eliminar) cuentas de manera proactiva (Tech Coalition, 2023). La plataforma ThreatExchange de Meta también permite a los participantes compartir y acceder a señales de una manera estructurada, y su documentación pública incluye numerosas señales potenciales, algunas de las cuales pueden ser muy granulares, como información sobre la huella del navegador de un usuario (“cadena de agente de usuario”), la dirección IP, así como latitud y longitud asociadas, o el nombre y la dirección revelados en una búsqueda “Quién es” para un dominio web asociado con una cuenta (Meta, 2024b).

Las plataformas de transmisión en vivo pueden utilizar señales independientemente de si son parte de dichos esfuerzos de intercambio de la industria o no. Por ejemplo, si los patrones de comportamiento en torno a una determinada transmisión son altamente inusuales (por ejemplo, como lo expresó un experto, la transmisión fue iniciada por una cuenta nueva y de repente tiene una gran audiencia que proviene toda de un enlace externo, con cuentas en su mayoría nuevas en el chat, la mayoría de las cuales están vinculadas a direcciones IP asociadas con servidores VPN y tienen correos

Como lo expresó un participante, la idea es que las empresas participen más activamente en reducir la capacidad de su plataforma para ser utilizada para la difusión de material de abuso sexual infantil, no solo participando en Detectar y Denunciar, sino también, aspiracionalmente, hacia un modelo de confianza y seguridad de Predecir e Interceptar.

electrónicos vinculados a proveedores de correo electrónico anónimos o que preservan la privacidad), estos factores pueden ayudar a categorizar una transmisión como de alto riesgo para una revisión rápida del moderador. Algunas empresas afirmaron que los enfoques de señales ya les están ayudando a tomar decisiones más informadas sobre la moderación de ciertas cuentas y, en general, son útiles desde una perspectiva de confianza y seguridad antes de que pueda realizarse una revisión humana. Las señales abren toda una gama de posibilidades en la caja de herramientas de moderación, que pueden ir más allá de simplemente eliminar contenido: si cierta actividad del usuario se califica como potencialmente riesgosa, las empresas pueden implementar fricción en las transmisiones en curso, incluida la reducción del ancho de banda en la transmisión, el aumento de la latencia en el chat o incluso el cierre de la sesión de los miembros de la audiencia de la transmisión para que necesiten volver a iniciar sesión mientras un moderador humano realiza una investigación.

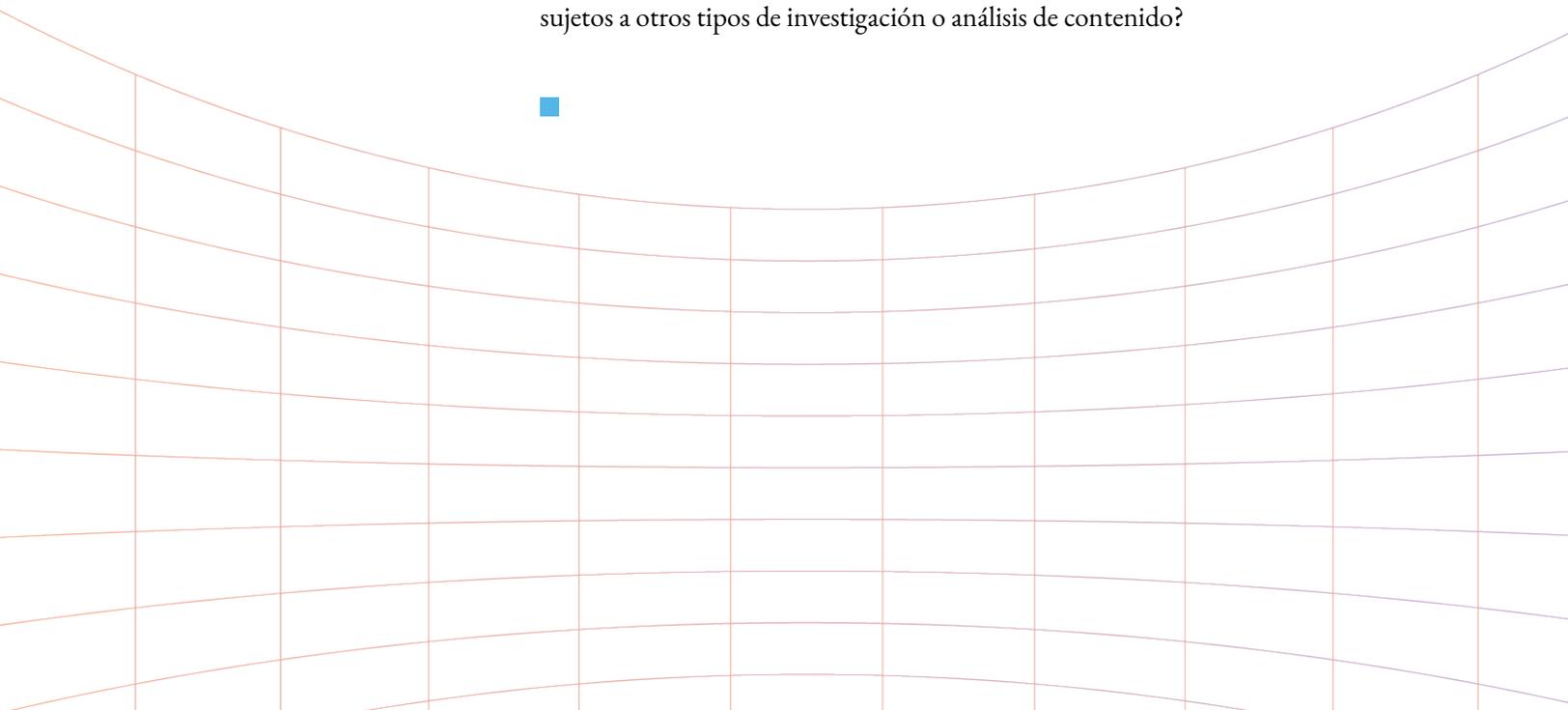
Se trata de una práctica industrial emergente que aún no ha sido objeto de un escrutinio exhaustivo por parte del sector académico, la sociedad civil o el periodismo. Tanto las empresas como los expertos sugieren que este es un enfoque muy prometedor en términos de eficacia, que permite a las empresas no solo intentar evitar de manera sólida que las cuentas de actores maliciosos compartan contenido (o que las cuentas eliminadas regresen a la plataforma), sino también potencialmente usarlo durante las transmisiones en combinación con otras técnicas para realizar intervenciones de moderación más precisas.

Como lo expresó un participante de la industria, las empresas más proactivas han dejado de ver sus esfuerzos por proteger la transmisión en vivo como un problema de moderación de contenido y, en cambio, parecen pensar en ello como un problema de “inteligencia de amenazas” vinculado al campo más amplio de la ciberseguridad y la prevención del fraude. Este giro es evidente en nuestras entrevistas no solo con plataformas de transmisión en vivo públicas, sino también con servicios de videollamadas, donde las empresas pueden decidir realizar intervenciones basadas no en comunicaciones privadas, sino más bien en metadatos circundantes. Como lo expresó un participante, la idea es que las empresas participen más activamente en reducir la capacidad de su plataforma para ser utilizada para la difusión de material de abuso sexual infantil, no solo participando en Detectar y Denunciar, sino también, aspiracionalmente, hacia un modelo de confianza y seguridad de Predecir e Interceptar. Investigaciones anteriores también señalaron que el énfasis en los metadatos en lugar del análisis del contenido real generado por el usuario era importante para las plataformas que ofrecen servicios de comunicación encriptados de extremo a extremo ([Kamara et al., 2021](#)).

Algunos participantes de la industria mencionaron que creían que las señales eran una herramienta poderosa para ayudar a las empresas a ser más selectivas en el uso de técnicas potencialmente más dañinas o invasivas de la privacidad, como las formas de análisis de contenido mencionadas en la sección anterior. En lugar de transcribir todas las transmisiones y ejecutar un modelo de análisis de texto en los resultados, en teoría, un enfoque de señales podría permitir a las empresas decidir qué transmisiones deben examinarse con mayor profundidad, reduciendo la posibilidad de falsos positivos aleatorios y minimizando el grado en que usuarios completamente aleatorios terminen atrapados en este tipo de redes automatizadas. Además, el análisis de señales podría abordar algunas de las deficiencias del análisis de contenido utilizando modelos lingüísticos grandes para examinar chats o transcripciones de audio en idiomas de “bajos recursos”.

Estos enfoques todavía están en sus etapas iniciales. Sin embargo, nuestras entrevistas también plantearon inquietudes respecto de que el uso masivo de señales ha evolucionado no solo para sugerir una investigación manual cuidadosa por parte de expertos (por ejemplo, los equipos de investigación de seguridad infantil de la plataforma), como era la práctica habitual en la industria, sino cada vez más para desarrollar modelos que toman decisiones automatizadas a escala y podrían llevar a consecuencias imprevistas. No está claro hasta qué punto las empresas utilizan señales y metadatos para desarrollar clasificadores.

En particular, se necesita más investigación para examinar el impacto de este conjunto de prácticas altamente opacas en desarrollo sobre poblaciones que pueden tener algunas características “sospechosas”. ¿Se está creando un perfil de usuarios preocupados por su privacidad (que utilizan VPN o proveedores de correo electrónico distintos de Gmail o Outlook) y se está viendo reducida su experiencia de usuario o se están eliminando sus cuentas por error? ¿Qué pasa con la amplia demografía de usuarios asociados a ciertos países de mayoría global que teóricamente también podrían ser identificados como sospechosos (debido a sus direcciones IP y características de ubicación) y, por lo tanto, sujetos a otros tipos de investigación o análisis de contenido?



III. Implicaciones Políticas de las Prácticas de Confianza y Seguridad Existentes

Estas tres amplias modalidades de intervenciones contra diversas formas de contenido sexual transmitido públicamente que involucra a menores tienen diferentes implicaciones, modos de falla e impactos en los derechos humanos de los usuarios de la plataforma en general y de los sobrevivientes en particular.

Limitaciones de las Intervenciones de Diseño Existentes

Las intervenciones basadas en diseño y estructura que establecen umbrales de popularidad antes de que una cuenta pueda transmitir en vivo probablemente eviten muchos tipos de abusos graves sin necesidad de lidiar con las deficiencias de privacidad y seguridad de las herramientas de verificación de edad. Aunque no son perfectos, los influencers populares cuyas cuentas han sido usurpadas o que infringen las normas de la comunidad con contenido sexual autogenerado tienen más probabilidades de ser denunciados por los espectadores y ser eliminados. Sin embargo, las plataformas que priorizan el streaming y que no tienen otras formas para que las cuentas generen seguidores orgánicos (como pueden hacerlo, por ejemplo, en TikTok o Instagram a través de su contenido que no es en vivo) no tienen este lujo y, por lo tanto, necesitan buscar otros enfoques.

Un enfoque que exige que solo los creadores de contenido verifiquen su edad (en lugar de solo los usuarios comunes/espectadores de transmisiones) puede parecer una medida de seguridad proporcional para algunos; después de todo, no poder transmitir no es el mismo tipo de barrera que no poder participar en absoluto en la esfera pública digital, pero implementar medidas de verificación de edad de manera efectiva y segura no solo es muy difícil de hacer de manera segura en general, sino que también es particularmente riesgoso cuando involucra los datos biométricos de los jóvenes (CNIL, 2022; Forland *et al.*, 2024). Estas tecnologías también pueden crear riesgos desproporcionadamente mayores para grupos específicos de niños, como aquellos con discapacidades (Bhatia & Aboulafia, 2024). Además, como señaló un participante, los enfoques que, por ejemplo, exigen una tarjeta de crédito registrada para reducir el riesgo de abuso podrían convertir estas formas de protección o señales en un bien de lujo. De hecho, debido a las cargas que imponen a todos los usuarios de una plataforma determinada, incluidos los impactos desproporcionados en las comunidades marginadas, los mandatos legales que exigen el uso de tecnologías de verificación de la edad pueden restringir o frenar la libertad de expresión protegida legalmente (Ruane *et al.*, 2024).

Ramificaciones del Análisis Predictivo de Videos en Transmisiones En Vivo: Sesgo, Bloqueo Excesivo, Calidad de los Datos y Fuentes de Datos

El régimen internacional de notificación de material de abuso sexual infantil detectado a la CyberTipline del NCMEC, lo cual es un requisito legal según la legislación de los Estados Unidos, y el consiguiente “registro de huellas digitales” de este contenido confirmado que debe compartirse con las empresas para sus esfuerzos de detección, puede ser imperfecto y susceptible de mejoras (*Grossman et al., 2024*). A pesar de sus problemas, es un sistema relativamente probado y establecido para identificar material de abuso sexual infantil conocido. El desafío para la prevención de material de abuso sexual infantil en el contexto de la transmisión en vivo es que la mayoría de los contenidos potencialmente infractores no pueden detectarse mediante esta técnica comprobada a lo largo del tiempo.

A medida que la industria avanza hacia otras técnicas para detectar material de abuso sexual infantil previamente desconocido —o se ve obligada a hacerlo por ley— aumenta considerablemente el potencial de falsos positivos y, por tanto, de efectos problemáticos e involuntarios en los usuarios ordinarios. Los sistemas de visión artificial diseñados para clasificar imágenes (en este caso, fotogramas de videos en vivo) son notoriamente propensos a la clasificación errónea, especialmente cuando se enfrentan a variaciones inesperadas en términos de raza y género (*Wang et al., 2022; Zhao et al., 2021*). Los sistemas de aprendizaje supervisado tienen dificultades fuera de su dominio y pueden fallar espectacularmente en el mundo real cuando enfrentan condiciones subóptimas (como videos de baja resolución, iluminación deficiente, figuras obstruidas o perfiles laterales de individuos). Algunas de las tecnologías centrales que sustentan los enfoques predictivos de la industria, como los modelos de estimación de edad, pueden exhibir sesgos raciales y de género preocupantes, así como problemas generales de desempeño que ha llevado a académicos críticos a cuestionar el uso de estos sistemas en absoluto (*Stardust et al., 2024*). A escala de plataforma, incluso los sistemas supuestamente precisos podrían producir muchos miles de falsos positivos a diario.

¿Cómo se obtienen realmente los datos de entrenamiento y se puede obtener el consentimiento antes de que los modelos (que podrían ser patentados y vendidos como un servicio con fines de lucro) sean entrenados con ellos? ¿Cómo pueden otras partes interesadas verificar el desempeño de los sistemas de la plataforma, incluidas las herramientas clave de análisis automatizado de contenido? En general, la predicción de “nuevo contenido” de material de abuso sexual infantil se enfrenta a un importante problema subyacente de evaluación comparativa.

Una investigación anterior del CDT ha explorado en profundidad los problemas futuros relacionados con la implementación de sistemas automatizados de análisis de contenido multimedia ([Shenkman *et al.*, 2021](#)). Aunque el enfoque de ese análisis era general y no específico para la prevención del material de abuso sexual infantil, las mismas preocupaciones relacionadas con estos métodos (cuestiones de robustez, calidad de los datos, falta de contexto, baja capacidad de medición y explicabilidad) son igualmente aplicables en este contexto. La mayoría de los ingenieros y gerentes de productos con orientación técnica de la industria con los que hablamos fueron sinceros respecto a las limitaciones de su tecnología, señalando que estas herramientas en general, son imperfectas y que ciertos niveles de sesgo son inherentes a los modelos de aprendizaje automático predictivo que buscan realizar clasificación de imágenes. Los proveedores, sin embargo, son más optimistas, viendo en esto una oportunidad comercial y buscando refinar su tecnología sobre la marcha una vez que ya haya sido desplegada.

Existen cuestiones adicionales relacionadas con la calidad de los datos, su suministro y la ética que son exclusivas del contexto de detección y prevención del material de abuso sexual infantil. ¿Cómo se obtienen realmente los datos de entrenamiento y se puede obtener el consentimiento antes de que los modelos (que podrían ser patentados y vendidos como un servicio con fines de lucro) sean entrenados con ellos? ¿Cómo pueden otras partes interesadas verificar el desempeño de los sistemas de la plataforma, incluidas las herramientas clave de análisis automatizado de contenido? En general, la predicción de “nuevo contenido” de material de abuso sexual infantil se enfrenta a un importante problema subyacente de evaluación comparativa. Actualmente no existen indicadores públicos de rendimiento que permitan a las empresas comprobar la precisión de sus sistemas o que permitan a los expertos, responsables políticos e investigadores comprender mejor su eficacia, así como el alcance de lo que es realmente posible ([Laranjeira da Silva *et al.*, 2022](#)).

Una cuestión clave es cómo entrenar modelos destinados a detectar material de abuso sexual infantil previamente desconocido, en particular considerando que el material de abuso sexual infantil es ilegal y, por lo tanto, no está disponible fácilmente para ser utilizado como datos de entrenamiento. En los últimos años, un número cada vez mayor de proveedores ha firmado contratos para desarrollar herramientas de investigación para las autoridades policiales. Un ejemplo es el proyecto multi-organización ARICA, financiado por la Comisión Europea ([ARICA, 2023](#)). Con contratos complejos, excepciones legales y acuerdos de intercambio de datos, nuestras entrevistas se refirieron a modelos específicamente entrenados para la clasificación de imágenes con conjuntos de datos gubernamentales de material de abuso sexual infantil confirmados. La mayoría de estos se utilizan para apoyar investigaciones policiales, pero algunos parecen comercializarse para uso comercial. Generalmente no existe documentación pública asociada con estos productos, pero nuestras conversaciones con los proveedores incluyeron afirmaciones recurrentes de que sus herramientas pueden predecir con precisión la edad de los jóvenes e identificar fotogramas de imágenes sexuales dentro de las transmisiones, proporcionar estimaciones precisas sin significativos sesgos raciales o de género.

Sin embargo, hay muchas razones para ser escépticos. Como planteó un ingeniero, los clasificadores de imágenes que buscan “nuevo” contenido de material de abuso sexual infantil están luchando una batalla difícil: Para funcionar bien a escala de plataforma, los sistemas deben ser “99,99999999% efectivos”, o de lo contrario estarán marcando erróneamente miles o incluso decenas de miles de piezas de contenido a diario.

Sin transparencia, información sobre pruebas o análisis comparativos estandarizados, es difícil evaluar la eficacia de estas afirmaciones. Sin embargo, hay muchas razones para ser escépticos. Como planteó un ingeniero, los clasificadores de imágenes que buscan “nuevo” contenido de material de abuso sexual infantil están luchando una batalla difícil: Para funcionar bien a escala de plataforma, los sistemas deben ser “99,99999999% efectivos”, o de lo contrario estarán marcando erróneamente miles o incluso decenas de miles de piezas de contenido a diario. Si no se integra en el tipo correcto de canal de aplicación, esto conducirá a eliminaciones erróneas de contenido y patrones imprevistos de supresión de contenido, potencialmente con los mayores efectos sobre las formas legítimas de expresión de índole sexual. Por un lado, se trata de un área de especial importancia y con importantes implicaciones para las víctimas de abuso sexual, y algunos entrevistados expresaron su preocupación de que ser demasiado cuidadosos con las repercusiones sobre los derechos humanos haría que las empresas desistieran en desarrollar tecnologías nuevas e innovadoras que pudieran ayudar a controlar ciertas formas de violencia física. Sin embargo, todavía existen preocupaciones de que las empresas no están supervisando o no quieren supervisar de manera sólida estos modelos y aplicarlos con el nivel de cuidado adecuado. Por ejemplo, una cita pública mostrada como respaldo por un importante proveedor de seguridad se jacta de que su cliente “[ni] se molesta en revisar el contenido que [la herramienta] señala: es así de bueno y consistente” ([ActiveFence, 2024](#)).

Uno de los principales desafíos está relacionado con la calidad de los datos y la capacidad de los modelos para lograr un dominio preciso en el problema específico que pretenden resolver. El enfoque clásico de resolución de problemas de informática para identificar contenido X en entornos abiertos implica recolectar grandes cantidades de contenido X y luego implementar diversos métodos para ajustar o entrenar un modelo de clasificación de imágenes basado en patrones encontrados en esos datos. En teoría, el modelo ideal de detección del “nuevo material de abuso sexual infantil” se entrenaría en algún tipo de conjunto de datos recopilados globalmente con contribuciones de las autoridades y las organizaciones aliadas para lograr un conjunto de datos de entrenamiento globalmente representativo; estos modelos luego se probarían en un sólido conjunto de datos de referencia multijurisdiccional lleno no solo de imágenes completamente no relacionadas (por ejemplo, de ImageNet u otros conjuntos de datos de clasificación de imágenes genéricos), sino también con muchas imágenes legítimas que representan a jóvenes y contenido sexual pero no contenido ilegal. Sin este tipo de colaboraciones u otras medidas relacionadas (auditorías de conjuntos de datos realizadas por expertos independientes con las habilidades y la capacitación necesarias, mejores prácticas de evaluación comparativa de modelos o ejercicios de trabajo en equipo realizados por terceros y otras formas de prueba y transparencia limitadas a las partes interesadas clave), es extremadamente difícil para los investigadores o los responsables políticos determinar si estos productos son algo más que un simple engaño de la IA.

Algunos participantes mencionaron el desafío que supone la obtención ética de datos. Actualmente no existe una práctica recomendada estándar para obtener el consentimiento para el uso posterior de los datos de las víctimas (por ejemplo, para entrenar modelos; consulte también [Laranjeira et al., 2022](#)). Los activistas de la sociedad civil entrevistados expresaron reservas sobre el uso de material sexual infantil para entrenar modelos y luego vender el acceso a los productos resultantes con fines de lucro, sin el consentimiento de las personas representadas. Las conversaciones de la industria también destacaron los desafíos de navegar en un entorno donde los datos disponibles son extremadamente sensibles y sujetos a restricciones legales, lo que dificulta su recopilación y manejo. Sin embargo, algunos proveedores nos sugirieron que no solo obtienen acceso legal a material de abuso sexual infantil sensible por parte de las autoridades o de sus socios, sino que también lo buscan activamente en foros ilícitos y en “las partes sucias de Internet”, presentando eso como su principal valor añadido, aunque estas afirmaciones son difíciles de verificar de forma independiente.

Problemas con el Uso Compartido de Señales: Privacidad, Opacidad y Pocos Recursos

Los expertos de la industria y el mundo académico son optimistas sobre el movimiento en curso hacia el uso de señales, alejándose del análisis de las transmisiones en vivo a través de una serie de diferentes tipos de plataformas. Según los promotores, este desarrollo — que todavía está en sus primeras etapas— no solo puede reducir la posibilidad de falsas alertas y eliminaciones erróneas de contenido realizadas por sistemas de análisis de medios automatizados imperfectos, sino que también debería ser una forma mucho más eficaz de bloquear e investigar el número relativamente bajo de actores maliciosos motivados.

El problema, sin embargo, es que la forma en que se utilizan las señales es opaca y puede conducir a sesgos y fallos difíciles de detectar tanto internamente (dentro de las empresas) como externamente (por parte de los miembros del público potencialmente afectados). Algunas señales potenciales son mucho menos concretas y menos auditables que las firmas hash MD5, PDQ u otras. Por ejemplo, la lista del proyecto Lantern incluye “palabras clave utilizadas para hacer grooming”, un concepto inherentemente difuso ([Tech Coalition, 2023](#)). Además, hay preguntas sobre el grado en que las diferentes empresas pueden comprender adecuadamente las técnicas de producción de datos y entrenamiento de modelos utilizadas para desarrollar estos clasificadores. La opacidad en torno al uso de señales, particularmente en cómo se utilizan para predecir actores o comportamientos maliciosos, ya sea en sistemas automatizados o no, significa que es difícil verificar de forma independiente su eficacia fuera de las declaraciones de la industria.

Más allá del contexto de transmisión en vivo, el uso general de tales medidas abiertas a través de diferentes áreas de la plataforma (por ejemplo, mensajes directos no cifrados) plantea un importante problema potencial de libertad de expresión si conduce a intervenciones automatizadas contra los usuarios o incluso identifica ciertos grupos para una investigación en profundidad.

El uso de señales también podría tener implicaciones importantes para la privacidad. Los activistas defensores de las libertades civiles en la UE ya se están movilizando contra el uso de herramientas de análisis de texto orientadas al contacto por parte de las plataformas. Patrick Breyer, miembro del Partido Pirata del Parlamento Europeo, presentó una demanda a principios de 2024 ante un tribunal alemán bajo la sospecha de que dicho “grooming” estaba relacionado con clasificadores de texto que se estaban desplegando en sus chats de Facebook Messenger (Breyer, 2024). El proyecto de reglamento de la Comisión Europea sobre la prevención del abuso sexual infantil ha dado lugar a un amplio debate sobre las tecnologías predictivas “anti-grooming” y, asimismo, ha estado sumido en la controversia y una amplia organización contra la propuesta por parte de la sociedad civil (EDRi, 2023). Si se utilizan enfoques orientados a las señales en áreas no cifradas que, no obstante, vienen con alguna expectativa de privacidad (como mensajes privados, como afirma Breyer), existe una preocupación adicional de que las inocuas charlas cotidianas de las personas queden atrapadas por error en este tipo de redes automatizadas.

Más allá del contexto de transmisión en vivo, el uso general de tales medidas abiertas a través de diferentes áreas de la plataforma (por ejemplo, mensajes directos no cifrados) plantea un importante problema potencial de libertad de expresión si conduce a intervenciones automatizadas contra los usuarios o incluso identifica ciertos grupos para una investigación en profundidad. En Estados Unidos, por ejemplo, la Ley REPORT ahora exige que las plataformas denuncien los casos de “incitación”, lo que lleva a lo que Riana Pfefferkorn ha descrito como “mayores incentivos para denunciar en exceso” y a la posibilidad de que “la charla inocente en línea (por ejemplo, el coqueteo entre dos adolescentes o un usuario que cita la letra de una canción sobre proxenetismo) se denuncie primero al NCMEC y de allí a la policía” (Pfefferkorn, 2023).

Con iniciativas que intentan permitir el intercambio de señales en toda la industria, estos problemas podrían exacerbarse junto con la introducción de nuevas preocupaciones. Por ejemplo, errores, sesgos y otros problemas relacionados con los datos (o decisiones tomadas sobre el ajuste de parámetros, etc.) que conducen a resultados problemáticos que podrían, tal vez sin que las empresas del mismo consorcio se den cuenta, proliferar en las plataformas, lo que generaría efectos más amplios en todo el ecosistema de la plataforma e integraría cuestiones relacionadas con la explicabilidad y la opacidad del sistema aún más profundamente en los flujos de trabajo de confianza y seguridad de la industria.

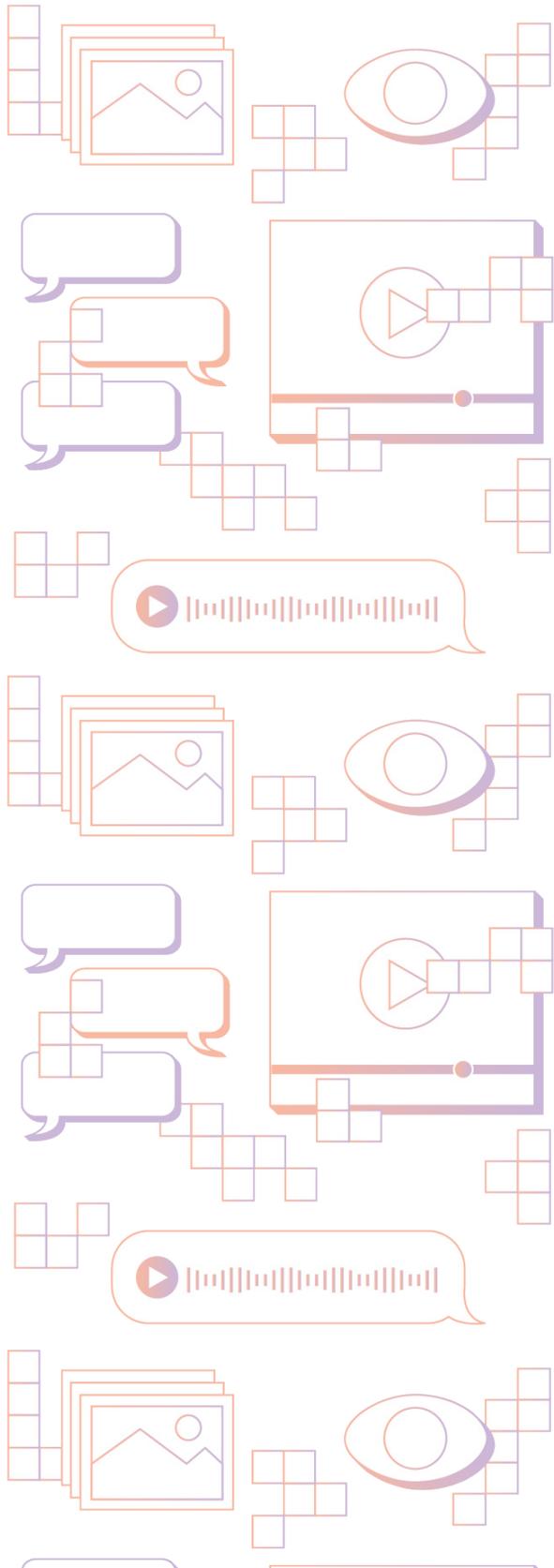
Consortios como Lantern plantean desafíos similares a los de sistemas como la base de datos hash del Foro Global de Internet para la Lucha contra el Terrorismo (GIFTC, por sus siglas en inglés). Por ejemplo, si la cuenta de alguien es hackeada y luego utilizada para difundir material de abuso sexual infantil y luego los indicadores de su cuenta (nombre de usuario, dirección IP, otros metadatos asociados) se comparten con otras empresas, es posible que los efectos de las credenciales robadas o filtradas en una plataforma tengan un efecto dominó importante y difícil de remediar que lleve a que se bloqueen también en muchos otros servicios. Al igual que con las preocupaciones de larga data sobre los hashes y el GIFTC, las cuestiones clave de rendición de cuentas se refieren a la forma en que el material se incorpora al consorcio, la forma en que es supervisado o auditado por expertos, y las estructuras más amplias de recursos existentes (Llansó, 2020).

Las preocupaciones en materia de privacidad pueden exacerbarse en la medida en que los datos personales son aportados y compartidos entre empresas con este tipo de marcos. Por ejemplo, una combinación de diferentes metadatos puede utilizarse para inferir información personal o sensible, incluidas las identidades (Kamara *et al.*, 2021). Además, aunque las empresas implementan medidas para evitar clasificaciones erróneas, aún es posible que ocurran y a menudo es difícil estar al tanto de ellas hasta que alguien tenga que buscar un recurso.

Algunas empresas nos hablaron sobre las formas en que utilizan los procesos de “apelaciones” para permitir que los usuarios eliminados erróneamente intenten restablecer sus cuentas, y las grandes oscilaciones en sus métricas de apelación de eliminación de cuentas les indican que sus modelos basados en señales se han vuelto demasiado sensibles y tal vez están tomando demasiada actividad legítima en sus esfuerzos por eliminar actividad spam, fraudulenta o potencialmente peligrosa que prolifera material de abuso sexual infantil. Por lo tanto, los procesos de apelación bien estructurados son cruciales para brindar a los usuarios recursos y la capacidad de impugnar decisiones de moderación erróneas, pero no deberían ser la única protección que las empresas implementen cuando confían en este tipo de técnicas de confianza y seguridad orientadas a las señales.



Conclusión



Las estrategias y sistemas específicos de confianza y seguridad que la industria de plataformas está desarrollando e implementando actualmente para prevenir de manera más efectiva la difusión de material de abuso sexual infantil a través de superficies de video en tiempo real están evolucionando rápida y continuamente. Se trata de un panorama complejo con poca documentación pública sobre las mejores prácticas. También se encuentra en un momento crítico, ya que el creciente interés internacional en las intervenciones políticas para abordar la seguridad infantil y los daños en línea también ha llevado a una proliferación de proveedores de “tecnología de seguridad” y otros grupos de terceros que impulsan una amplia gama de soluciones técnicas y tecnologías hechas a la medida de análisis de contenido automatizado contra la explotación y el abuso sexual infantil.

Los proveedores de servicios en línea que operan áreas de video en tiempo real de todo tipo enfrentan una fuerte presión para intensificar sus esfuerzos para limitar la proliferación de la explotación y el abuso sexual infantil y los daños relacionados. Lo que está en juego es especialmente importante dado el actual movimiento transnacional de regulación de plataformas orientadas a la seguridad infantil y los proyectos de ley sobre “daños en línea”, los esfuerzos por controlar más estrictamente el acceso de los jóvenes a cierta información y servicios en línea (Marwick et al., 2024; Witting, 2019), y renovados esfuerzos de “solucionismo tecno-legal” aplicados a través de cambios de políticas y diseños supuestamente “aptos para niños” (Ángel & Boyd, 2024). Las disputas políticas centradas en ciertas plataformas adyacentes a la transmisión en vivo, especialmente los sitios de pornografía y los servicios de videollamadas, están implicando cada vez más temas más amplios como el cifrado de extremo a extremo y la verificación de usuario/edad, que son parte de debates extremadamente importantes en torno a la ciberseguridad, la privacidad y la autonomía digital no solo para los jóvenes, sino también para los adultos (Child Rights International Network & defenddigitalme, 2023; Forland et al., 2024; McKee & Lumby, 2022).

Un camino para avanzar esta conversación implicará un compromiso continuo entre empresas, proveedores, sociedad civil y expertos académicos, idealmente con más experimentación abierta y un compromiso con los sistemas técnicos reales que proponen estos actores. Hay varias áreas en las que se puede avanzar en este sentido:

Se necesita una mayor transparencia para ayudar a mejorar los esfuerzos para abordar la explotación y el abuso sexual infantil en las plataformas de transmisión en vivo.

Es urgente contar con mecanismos para mejores prácticas de evaluación comparativa para modelos de clasificación predictiva de imágenes, así como potencialmente para modelos de elaboración de perfiles de “riesgo del usuario” orientados a “señales” de alto riesgo. Sin embargo, en la actualidad los proveedores tienen pocos incentivos para someter sus productos a formas tan abiertas de revisión, dado que un desempeño deficiente en los parámetros estándar de la industria tendría un impacto importante en su capacidad para exponerse a los clientes. Algunas de las mejores prácticas estándar de la investigación de vanguardia del aprendizaje automático aún no se han incorporado a este sector cada vez más politizado y sensible, pero está claro que al menos algunas formas simples de transparencia de conjuntos de datos y modelos (por ejemplo, tarjetas de modelos y hojas de datos que brinden información sobre cómo se entrenaron los modelos ([Crisan et al., 2022](#); [Mitchell et al., 2019](#))) podría incorporarse por empresas y proveedores. Si esto no se hace de manera totalmente pública, entonces se debe hacer de manera restringida y al alcance de las partes interesadas clave en la política, la sociedad civil y la industria. También podría ser útil compartir investigaciones patrocinadas o realizadas por proveedores. En varias entrevistas, nos enteramos de informes y documentos técnicos no públicos que están disponibles para clientes potenciales pero no para el público. Aunque existen marcos de transparencia voluntarios, véase por ejemplo ([Tech Coalition, 2022](#)), los desafíos que enfrentamos al realizar esta investigación y tratar de aprender cómo funcionan los sistemas de confianza y seguridad de la industria sugieren que estos marcos no son suficientes.

Lo que está en juego es especialmente importante dado el actual movimiento transnacional de regulación de plataformas orientadas a la seguridad infantil y los proyectos de ley sobre “daños en línea”, los esfuerzos por controlar más estrictamente el acceso de los jóvenes a cierta información y servicios en línea.

Los proveedores y las plataformas de transmisión en vivo deben ser explícitos acerca de las limitaciones de los enfoques automatizados para detectar y abordar la explotación y el abuso sexual infantil y desarrollar sus sistemas de confianza y seguridad en consecuencia.

Las empresas, conscientes de las limitaciones técnicas inherentes a los modelos de aprendizaje automático predictivo, deben seguir prestando atención a la forma en que se despliegan estos sistemas, estableciendo salvaguardias adecuadas para intentar mitigar los sesgos y la supresión de la expresión legítima. El uso de modelos predictivos para dar prioridad a transmisiones para una revisión rápida parece ser una medida razonable, que ayuda a las empresas a identificar cuándo involucrar a un revisor humano y, con suerte, les permite tomar decisiones matizadas según el contexto. Estos revisores deben ser moderadores expertos con una remuneración justa, y recibir capacitación especializada y apoyo psicológico para ayudarlos a lidiar con la posible exposición a material de abuso sexual infantil y contenido relacionado con material de abuso sexual infantil.

Centrarse en intervenciones de diseño que empoderen a los usuarios, incluidos los menores.

Aunque muchas de las intervenciones de diseño de las principales plataformas de transmisión en vivo se centran en la verificación y autenticación de la identidad, algunos esfuerzos incluyen herramientas para la moderación de los streamers y la comunidad. Las necesidades de los streamers de protegerse de ser blanco de ataques o de ser utilizados para distribuir la explotación y el abuso sexual infantil merecen mayor atención cuando se trata de soluciones basadas en el diseño. Por ejemplo, un enfoque basado en el diseño que no se planteó en nuestras conversaciones con la industria es proporcionar a los usuarios, en particular a los menores, el conjunto adecuado de herramientas y mecanismos de denuncia para ayudarlos a protegerse. La denuncia puede ser una herramienta importante para que los niños aborden problemas de la explotación y el abuso sexual infantil, como el grooming en línea (Kennedy et al., 2024). Investigaciones anteriores sugieren que, en el caso de las plataformas de mensajería directa, tener la capacidad de rastrear la respuesta de la plataforma a la denuncia de un usuario es particularmente importante (Luria, 2023). Esto también es relevante para las plataformas de transmisión en vivo, y herramientas como estas podrían personalizarse para tener en cuenta los perfiles de riesgo únicos de grupos específicos de menores (Luria, 2023).

Los modelos de gobernanza de múltiples partes interesadas pueden mejorar la rendición de cuentas de los enfoques para abordar la explotación y el abuso sexual infantil en una transmisión en vivo.

Los marcos de mejores prácticas en torno a la implementación de estos sistemas podrían desarrollarse no solo a través del trabajo continuo de organizaciones como la Tech Coalition, sino también a través de la participación crítica de múltiples partes interesadas en foros que no solo involucren a organizaciones de seguridad infantil, sino también a organizaciones que participan activamente en un conjunto más amplio de derechos digitales y libertades civiles. Aislar las conversaciones entre distintos grupos de partes interesadas en el gobierno y la sociedad civil no es un modelo sostenible a largo plazo para estos debates cruciales sobre políticas. De hecho, los modelos de múltiples partes interesadas aplicados a los mecanismos de gobernanza en torno al contenido terrorista, por ejemplo, se han beneficiado de este enfoque y se puede mejorar con una mayor rendición de cuentas (Bhatia, 2024). Por ejemplo, pueden permitir que los actores relevantes aporten información para el diseño de auditorías y otras formas de evaluaciones externas. Esto puede ser especialmente relevante para las medidas relacionadas con las señales. El lanzamiento inicial del proyecto Lantern por parte de Tech Coalition incluyó la noticia de que habían encargado una evaluación externa del impacto de la iniciativa en los derechos humanos (Tech Coalition, 2023) y la participación de firmas de auditoría especializadas e informes de terceros debería ser bienvenida en el futuro, dadas las grandes implicaciones del área en juego.

En general, nos encontramos en un momento clave para el futuro a la hora de abordar la explotación y el abuso sexual infantil, específicamente en las plataformas de transmisión en vivo. Abordar este problema es de vital importancia dado el impacto en los niños, los padres y sus comunidades, ya que se trata de un área de gobernanza de plataformas de enorme importancia y trascendencia. Es comprensible que tanto los proveedores como la industria estén ansiosos por demostrar que están desarrollando nuevas herramientas innovadoras para manejar las demandas de las partes interesadas y que están tomando en serio el área general de la seguridad infantil y el abuso sexual infantil, pero una implementación deficiente (o un diseño deficiente de confianza y seguridad, con sistemas que son fundamentalmente defectuosos) disminuirá, en lugar de aumentar, la confianza de los responsables políticos y del público en la confianza y la seguridad de las plataformas en el largo plazo. Por lo tanto, todas las partes interesadas involucradas deberían tener interés en garantizar que las prácticas emergentes de la industria se implementen de una manera cuidadosa y responsable, basada en una evaluación realista de las posibles compensaciones, limitaciones tecnológicas y efectos colaterales que estas diferentes intervenciones podrían tener.



Referencias

- 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., & Aboulaflia, 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/S5]9-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/8]JB9-2]XB]
- 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.io*. 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

