La cooperación entre plataformas como mecanismo para abordar la moderación de contenido en línea

Revista Latinoamericana de Economía y Sociedad Digital

Issue Especial 2

Autores: Federica Tortorella Casilla

DOI: 10.53857/RLESD.04.2023.04

Publicado: 10 marzo, 2024

Recibido: 6 diciembre, 2023

Cita sugerida: Tortorella Casilla, F. (2023). La cooperación entre plataformas como mecanismo para abordar la moderación de contenido en línea, Revista Latinoamericana de Economía y Sociedad Digital(4)

Tipo: Ensayo

Resumen

Moderar contenido en línea desde un enfoque de respeto a los derechos y libertades es una tarea que aún sigue pendiente. En este escenario, el concepto de cooperación entre plataformas ha sido reivindicado en múltiples ocasiones; al momento, este modelo ha sido aplicado para contrarrestar la publicación de contenido sexual infantil explícito y de tipo terrorista o extremista. Además de revisar algunas tendencias actuales y definiciones relacionadas con el tema de análisis, el objetivo de esta investigación es procurar mayor entendimiento del concepto de cooperación entre plataformas y de su actual aplicación. A partir del estudio del caso del Global Internet Forum to Counter Terrorism, principalmente, este trabajo invita a analizar los principios y mecanismos de dicho foro, para entender si podrían ser de utilidad en la mejora de los esfuerzos encaminados a contrarrestar el contenido dañino en línea. La investigación finaliza con una propuesta para crear un modelo de moderación de contenido en línea, que esté orientado a los derechos humanos, promoviendo la noción de cooperación entre plataformas, como la aplicada por el organismo aquí tratado, e incluyendo conceptos que surgen de un planteamiento del Global Partners Digital, así como el uso del modelo de múltiples partes interesadas que caracteriza a

iniciativas como ICANN.

Abstract

Moderating online content from an approach of respect for rights and freedoms is a task that is still pending. In this scenario, the concept of cooperation between platforms has been vindicated on multiple occasions; at the moment, this model has been applied to counteract the publication of explicit child sexual content and terrorist or extremist content. In addition to reviewing some current trends and definitions related to the topic of analysis, the objective of this research is to seek a better understanding of the concept of cross-platform cooperation and its current application. Based primarily on the case study of the Global Internet Forum to Counter Terrorism, this paper invites an analysis of the principles and mechanisms of the Global Internet Forum to Counter Terrorism to understand whether they could be useful in enhancing efforts to counter harmful online content. The research concludes with a proposal to create a human rights-oriented model of online content moderation, promoting the notion of cross-platform cooperation, as applied by the organization discussed here, and including concepts arising from a Global Partners Digital approach, as well as the use of the multi-stakeholder model that characterizes initiatives such as ICANN.

Resumo

A moderação de conteúdo on-line a partir de uma abordagem que respeite os direitos e as liberdades é uma tarefa que ainda está pendente. Nesse cenário, o conceito de cooperação entre plataformas tem sido justificado em várias ocasiões; no momento, esse modelo tem sido aplicado para combater a publicação de conteúdo sexual explícito para crianças e conteúdo terrorista ou extremista. Além de revisar algumas tendências e definições atuais relacionadas ao tópico de análise, o objetivo desta pesquisa é buscar maior compreensão do conceito de cooperação entre plataformas e sua aplicação atual. Baseando-se principalmente no estudo de caso do Fórum Global da Internet de Combate ao Terrorismo, este documento convida a uma análise dos princípios e mecanismos do Fórum Global da Internet de Combate ao Terrorismo para entender se eles poderiam ser úteis para aumentar os esforços de combate ao conteúdo nocivo on-line. A pesquisa conclui com uma proposta de criação de um modelo de moderação de conteúdo on-line orientado para os direitos humanos, promovendo a noção de cooperação entre plataformas, conforme aplicada pelo órgão discutido aqui, e incluindo conceitos decorrentes de uma abordagem de Parceiros Globais Digitais, bem como o uso do modelo de múltiplas partes interessadas que caracteriza iniciativas como a ICANN.

Palabras clave: moderación de contenido en línea, cooperación entre plataformas digitales.

1. Introducción

A lo largo de los años, la moderación de contenido en línea ha sido abordada desde diferentes perspectivas. Al respecto, pueden contarse varias iniciativas, por ejemplo, la implementación de marcos regulatorios, la adopción de acuerdos y principios, así como el establecimiento de mecanismos de colaboración entre plataformas digitales para detectar y eliminar contenido considerado prohibido y/o ilícito. Este último tipo de iniciativa será analizado en la presente investigación, tomando como caso de estudio la experiencia del Global Internet Forum to Counter Terrorism (GIFCT).

Desde el año 2017, algunas de las grandes plataformas cooperan entre sí para contrarrestar el uso de sus espacios para la publicación de contenido dañino de tipo extremista. En este escenario, el Global Internet Forum to Counter Terrorism nació como una organización compuesta por varias de las grandes empresas tecnológicas (big tech) entre las que destacan Twitter –hoy X–, Microsoft, Facebook y YouTube. Esta iniciativa consiste en la creación de una base de datos de *hashes* alimentada por cada miembro, la cual es utilizada para detectar y remover contenido dañino que viole sus políticas. Con el Llamamiento de Christchurch, en 2019, el GIFCT se unió a la visión impulsada por los gobiernos de Francia y de Nueva Zelanda para luchar contra el contenido terrorista y extremista en línea.

El análisis de la cooperación entre plataformas, planteada por el GIFCT, nos lleva a las considerar las siguientes preguntas: ¿es posible y/o deseable replicar este modelo para moderar otro tipo de contenido dañino?, ¿cuáles son las ventajas y desventajas del modelo tal como lo conocemos hasta ahora?, ¿existen otras iniciativas o propuestas que aborden los desafíos de la moderación de contenido?, ¿alguna de estas propuestas contempla la cooperación entre plataformas? En relación con las iniciativas y propuestas que ya existen, ¿qué particularidades o características deberían ser implementadas para moderar contenido en línea con un enfoque basado en los derechos humanos?

En la primera parte de esta investigación se definen conceptos fundamentales para el análisis del caso de estudio, tales como libertad de expresión, moderación de contenido en línea y cooperación entre plataformas, para luego examinar la iniciativa GIFCT desde sus inicios y explicar en qué se basa su labor de *prevención, respuesta e investigación* y cómo la organización y sus proyectos han ido evolucionando hasta ahora. En la segunda parte se enuncian las ventajas y desventajas de este modelo, a partir de nociones como factibilidad y representatividad. Antes de finalizar, se exploran los términos plasmados en el documento técnico (*white paper)* del Global Partners Digital, para poder avanzar una propuesta de

cooperación entre plataformas que promueva, entre otros, transparencia, representatividad y un marco más apegado a los derechos humanos.

2. Libertad de expresión, moderación de contenido en línea y cooperación entre plataformas: una mirada general

2.1 Libertad de expresión

Libertad de expresión es un concepto mencionado a menudo en diversos ámbitos y escenarios de la actual vida cotidiana; sin embargo, ¿qué es con exactitud y cuál es su alcance? Para responder, podríamos mencionar tantas fuentes que sería necesario dedicar esta investigación por completo a ello, pero aquí nos limitaremos a citar algunos de los múltiples tratados internacionales en los que la libertad de expresión es definida y que ofrecen garantías para su ejercicio.

En la Carta de los Derechos Fundamentales de la Unión Europea (2000) se determina que la libertad de expresión y de información "comprende la libertad de opinión y la libertad de recibir o comunicar informaciones o ideas sin que pueda haber injerencia de autoridades públicas y sin consideración de fronteras" (artículo 11). Respecto al contexto de las Américas, en la Declaración Americana sobre los Derechos y Deberes del Hombre (1948) se habla de la "libertad de investigación, de opinión y de expresión y difusión del pensamiento por cualquier medio" (artículo IV), y en la Convención Americana sobre Derechos Humanos (1969) se aborda la libertad de pensamiento y de expresión en estos términos: "toda persona tiene derecho a la libertad de pensamiento y de expresión. Este derecho comprende la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística, o por cualquier otro procedimiento de su elección" (artículo 13). Por último, en la Declaración Universal de Derechos Humanos (1948), la libertad de opinión y de expresión es considerada en el sentido de que "este derecho incluye el de no ser molestado a causa de sus opiniones, el de investigar y recibir informaciones y opiniones, y el de difundirlas, sin limitación de fronteras, por cualquier medio de expresión" (artículo 19).

Al igual que otros derechos y libertades, la libertad de expresión no está exenta de análisis como tampoco del reconocimiento y la necesidad de ser garantizada en las plataformas digitales. La llegada de la internet y el surgimiento de este tipo de plataformas han generado interés y urgencia para dar respuesta a problemáticas relacionadas con cómo

actuar para seguir protegiendo y garantizando este derecho frente a los desafíos que el ecosistema digital conlleva. En particular, y teniendo en cuenta que tal ecosistema traspasa las barreras y fronteras físicas, de esta premisa surge otra inquietud respecto a qué mecanismos son los más adecuados para evitar la violación de las libertades, considerando la diversidad cultural, social y política en el mundo, que no deja de reflejarse en los espacios en línea.

2.2 Moderación de contenidos

Junto a la libertad de expresión, la moderación de contenido en línea es otro concepto que ha acaparado la atención de distintos actores dada la estrecha relación que hay entre ambos términos. La complejidad de esta noción propicia múltiples definiciones; así, es contemplada como "[el conjunto de] actividades realizadas por los prestadores de servicios intermediarios destinadas a detectar, identificar y actuar contra contenidos ilícitos o información incompatible con sus condiciones, que los destinatarios del servicio hayan proporcionado" (Santisteban, 2022, p. 161). La organización Artículo 19 establece que dicha moderación "incluye los diferentes conjuntos de medidas y herramientas que las plataformas usan para hacer frente a los contenidos ilegales y aplicar sus normas comunitarias a los contenidos generados por los usuarios en sus servicios", de modo que, en general, "implica la señalización por parte de los usuarios de confianza o 'filtros', la eliminación, etiquetado, desclasificación o desmonetización de contenidos, o la desactivación de determinadas funciones" (Belli et al., 2021). Además de su acción correctiva, al ocuparse del manejo de contenido dañino en línea, este tipo de moderación previene la creación de ambientes hostiles, apoyando, en cambio, la comunicación positiva en línea, con lo cual disminuyen las agresiones y conductas antisociales.

Hablar de moderación de contenido en línea implica abordar ciertos conceptos y sus diferencias; por ejemplo, es importante diferenciar entre contenido considerado ilícito en función de marcos legislativos y regulatorios locales y contenido que infringe las políticas de las plataformas digitales. En múltiples instancias se han ido reconociendo algunos requerimientos y limitaciones por parte de los Estados y las plataformas en torno a la remoción de contenido de tipo extremista, que genere desinformación y discursos de odio, entre otros, bajo la premisa de que se trata de contenido que violenta los derechos de los internautas. En este sentido, las declaraciones conjuntas *Sobre Libertad de Expresión y Lucha contra el Extremismo Violento* y *Sobre Libertad de Expresión y "Fake News", Desinformación y Propaganda*, de 2016 y 2017, respectivamente, exigen a los Estados no someter a las plataformas digitales mediante órdenes para remover o restringir contenido, a menos de que este se trate de tipologías identificadas y prohibidas por estándares

internacionales, contrarrestando, así, la censura; mientras que a las plataformas se les reclama fácil acceso y entendimiento de sus políticas y lineamientos, los cuales deben ser formulados acorde a los parámetros de los derechos humanos.

Asimismo, en los Principios Rectores sobre Empresas y Derechos Humanos, de las Naciones Unidas, se establece que los Estados deberán adoptar medidas contra la violación de los derechos humanos, así como requerir a las empresas que desarrollen sus operaciones siguiendo el mismo precepto. Las empresas –en las que se incluye a las plataformas digitales– tienen la obligación de respetar los derechos humanos, siendo estos "entendidos, como mínimo, como los expresados en la Carta Internacional de Derechos Humanos y los principios relativos a los derechos fundamentales establecidos en la Declaración de la Organización Internacional del Trabajo relativa a los principios y derechos fundamentales en el trabajo" (Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos, 2011).

2.3 Responsabilidad de intermediarios

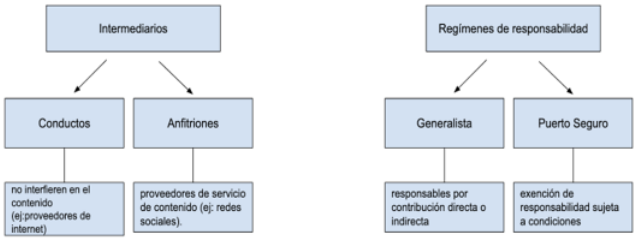
En ese mismo orden, es también oportuno repasar los regímenes de responsabilidad de los intermediarios y su relación con la moderación de contenido en línea. Si bien es cierto que esta última es necesaria para contrarrestar la presencia y propagación de contenido ilícito en línea, también podría generar responsabilidad por parte de los intermediarios.

Cuando hablamos de intermediarios, nos referimos a "una entidad que presta servicios que permiten el uso de Internet por parte de los usuarios" (Association for Progressive Communications, 2014). Los intermediarios pueden ser divididos en dos grandes categorías, siendo los "conductos" o *conduits* aquellos que no interfieren en el contenido transmitido, excepto para el almacenamiento, mientras que los "anfitriones" o *hosts* son proveedores de servicios de contenido. En la primera categoría se pueden incluir, por ejemplo, a los operadores de red móvil y proveedores de servicios de internet, y en la segunda, a las redes sociales y blogs.

Los regímenes de responsabilidad empleados actualmente son: "generalista" o *generalist* y "puerto seguro" o *safe harbour*. En el primer caso, los intermediarios llegan a ser responsables de los contenidos, bien por contribución directa o indirecta; mientras que el modelo de "puerto seguro" prevé exención de responsabilidad sujeta a condiciones que pueden ser muy detalladas y estrictas o diseñadas para hacer frente a diferentes tipos de

responsabilidad y actividades en distintos ámbitos de la ley (Association for Progressive Communications, 2014).

Figura 1. Tipos de intermediarios y regímenes de responsabilidad



Fuente: elaboración propia.

En el análisis de la relación entre moderación de contenido en línea y responsabilidad de intermediarios es oportuno mencionar la sección 230 de la Ley de Decencia en las Comunicaciones o Communications Decency Act (CDA), al igual que los Principios de Manila sobre Responsabilidad de Intermediarios. La primera fue aprobada en Estados Unidos en 1996; en particular, en su sección 230, la CDA provee de inmunidad a los intermediarios por los contenidos que hayan generado sus usuarios, siendo estos últimos los únicos responsables, y protege a los intermediarios aun cuando moderen esos contenidos en sus plataformas. La promulgación de esta ley ha permitido el desarrollo de la internet tal como la conocemos ahora, es decir, con la CDA ha sido posible evitar la censura y proteger la libertad de expresión, en especial de los grupos más vulnerables, ya que su aplicación va desde las plataformas más pequeñas hasta las *big tech*, sin distinciones. Las únicas excepciones a este artículo son las reclamaciones federales de propiedad intelectual y leyes federales de tipo penal, como lo son SESTA (Stop Enabling Sex Traffickers Act) y FOSTA (Fight Online Sex Trafficking Act), las cuales, cabe resaltar, han sido ampliamente criticadas, entre otras razones, por vulnerar la libertad de expresión de trabajadores sexuales (*What is SESTA/FOSTA?*, 2020; Silva, 2022).

Por su parte, los Principios de Manila consisten en un esfuerzo de la sociedad civil que entre 2014 y 2015 acordó seis preceptos para “orientar a los gobiernos, la industria y sociedad civil en el desarrollo de las mejores prácticas relacionadas con la regulación de los contenidos en línea a través de intermediarios” (*The Manila Principles on Intermediary Liability*, 2015, p. 2). Con base en instrumentos internacionales sobre derechos humanos y

otros marcos legales, estos principios consideran necesarios, entre otros aspectos: la protección de los intermediarios ante contenido de terceros y las órdenes judiciales como mecanismo para formalizar la restricción de contenido; el respeto del debido proceso, y la transparencia y rendición de cuentas en la normativa, políticas y prácticas en la materia.

2.4 Distintos abordajes de las complejidades en la moderación de contenidos

La moderación de contenidos despierta muchos interrogantes, además, se caracteriza por cierta complejidad que llega a influir en las iniciativas presentadas por las partes interesadas y, por ende, en su éxito. Por ejemplo, desde el punto de vista regulatorio, los gobiernos han ido promulgando leyes como la NetzDG, en Alemania, y la Ley de Servicios Digitales, de la Unión Europea. No obstante, se ha visto que normas nacionales con frecuencia entran en conflicto o no siempre son aplicables de manera directa. Por otro lado, el trabajo de la sociedad civil ha permitido crear espacios y tener como resultado textos como las Recomendaciones de Access Now sobre Gobernanza de Contenido y los Principios de Santa Clara; sin embargo, este sector también ha estado denunciando las condiciones precarias de los moderadores de contenido y los sesgos reflejados por los algoritmos (Andreu, 2022).

Otros intentos de abordar la moderación de contenidos en línea han sido llevados a cabo mediante acuerdos de cooperación, entre ellos, el memorándum de entendimiento sobre derechos en las plataformas ante moderación de contenido, firmado en Argentina. Si bien es cierto que a la fecha algunas de estas aproximaciones han presentado dificultades en su implementación o eficiencia, es válido aclarar que esto no constituye un impedimento para poder contar en un futuro con iniciativas similares, mejor estructuradas y de mayor grado de madurez, acompañadas de otros mecanismos, como la cooperación entre plataformas.

Para los fines de esta investigación nos concentramos justo en el concepto de cooperación entre plataformas; las razones principales de ello radican en que tal cooperación es un mecanismo que, en cierto modo, involucra múltiples actores del ecosistema de internet, además de que en los últimos años ha experimentado un incremento en su demanda (Douek, 2020).

Un caso reciente de cooperación entre plataformas proviene de la Unesco, que en febrero del presente año, 2023, auspició la conferencia *Internet para la confianza*, evento en el que “se debatió un proyecto que incluyera directrices para regular las plataformas digitales, con

un enfoque de múltiples partes interesadas para salvaguardar la libertad de expresión y el acceso a la información digitales" (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 2023). Otras iniciativas basadas en la cooperación entre plataformas son:

- La Global Network Initiative, que reúne empresas de telecomunicación, sociedad civil, academia y plataformas digitales para conformar una alianza que garantice el respeto a la libertad de expresión y el derecho a la intimidad frente a las presiones de los gobiernos;

- El Plan de acción de la Comisión Europea para enfrentar la desinformación, el cual consta de una serie de iniciativas que incluyen directrices, herramientas a utilizar, programas de monitoreo, entre otros, y

- El Foro Global de Internet para Contrarrestar el Terrorismo o Global Internet Forum to Counter Terrorism (GIFCT), cuyos miembros trabajan en conjunto para la erradicación del contenido extremista en línea.

3. **Sobre Christchurch Call y el GIFCT**

En marzo de 2019 se produjo un episodio que marcó profundamente la moderación de contenido en línea: dos tiroteos masivos llevados a cabo por un supremacista blanco en las mezquitas de Christchurch, en Nueva Zelanda, fueron transmitidos en vivo, y la grabación, de unos 17 minutos, fue vista más de cuatro mil veces antes de ser eliminada. A raíz de lo ocurrido, los gobiernos de Francia y de Nueva Zelanda crearon el Christchurch Call (Llamamiento de Christchurch), con la intención de fomentar la cooperación entre gobiernos y plataformas digitales para contrarrestar el contenido terrorista y extremista en las redes. El Global Internet Forum to Counter Terrorism, que ya existía desde 2017 como plataforma de cooperación entre algunas *big tech* con el fin de prevenir la explotación de los medios digitales para fines terroristas y extremistas, se unió junto con los demás sectores al llamamiento (*call to action*) impulsado por la iniciativa. En septiembre 2019, las compañías fundadoras (Twitter, Microsoft, Facebook y YouTube) anunciaron la independización del GIFCT y cambios en el modelo de gobernanza; así, el Foro pasó de ser una agrupación de múltiples partes interesadas (*multistakeholder*) a una organización sin fines de lucro, con un director ejecutivo, un comité asesor independiente y una junta operativa.

El GIFCT se rige por tres pilares: prevención, respuesta y aprendizaje/investigación (*prevent, respond and learn*); en cuanto a la investigación, respalda a la Red Global sobre Extremismo y Tecnología o Global Network on Extremism and Technology (GNET) y cuenta con un marco de definiciones y principios. Desde 2020, los grupos de trabajo creados por el GIFCT se encargan de labores que afianzan sus tres pilares. Además, esta organización es socia y financiadora del Tech Against Humanity, una iniciativa que apoya la industria digital en sus esfuerzos por combatir la explotación de las redes para fines terroristas, promoviendo los derechos humanos. Como herramientas de prevención y respuesta, existe la base de datos de *hashes* y el Protocolo para Incidentes de Contenido o Content Incident Protocol (CIP).

La base de datos de *hashes,* utilizada por el GIFCT desde su creación, está conformada con la recopilación de contenido extremista que ha sido removido por las plataformas integrantes de dicha organización, para ser almacenado en forma de representación numérica no reversible (*perceptual hashes*). La existencia de esta base le permite a la membresía enterarse sobre el contenido eliminado por otros integrantes, verificar si este infringe sus términos y condiciones, y, de haber algo similar en sus espacios, poder removerlo. Es importante especificar que se toma como referencia la lista de grupos designados como terroristas por la ONU y que la base de datos se auxilia de un protocolo llamado Discrete Cosine Transform, encargado de convertir y formatear el contenido subido a la base misma justo para crear un *hash*. En este sentido, los fundadores del GIFCT adoptaron una taxonomía (Hash-Sharing Database Taxonomy), que fue ampliada a raíz de los acontecimientos de Christchurch y sometida a revisión en 2021; esta última dio como resultado, entre otras novedades, la inclusión del concepto de "proclamaciones por parte de los atacantes" (*attackers' manifesto*) a la lista de contenido dañino.

De forma simultánea, la base de datos apoya la implementación del CIP, el cual es activado exclusivamente en caso de presentarse cuatro condiciones:

1. La detección de "un evento terrorista, extremista violento o de violencia masiva del mundo real";
2. que el contenido haya sido transmitido en vivo o grabado;
3. que este represente asesinatos o intento de asesinato, y
4. que su distribución se realice en las plataformas de los miembros de GIFCT o de una manera tan amplia en línea, que la difusión resulte inevitable.

El uso más reciente del Protocolo fue registrado durante el tiroteo masivo en Memphis, Tennessee, en septiembre de 2022.

Con lo anterior, vemos cómo el GIFCT evolucionó hasta ser una organización sin fines de lucro y un referente en materia de moderación de contenido de tipo extremista. Según el reporte anual de 2022, el Foro cuenta con 22 plataformas digitales en calidad de miembros, y sigue fortaleciendo las labores enfocadas en sus tres pilares mediante actividades como: programas de mentorías para empresas digitales, ampliación de su taxonomía, participación en espacios de diálogo interinstitucionales y continuidad de las labores de los grupos de trabajo. Sin embargo, los desafíos con respecto a la moderación de contenido en línea no se limitan a la tipología de contenido dañino (*harmful content*).

4. Cooperación entre plataformas para moderar contenido: ventajas y desventajas

Analizando el GIFCT y su historia podemos observar que: *1*) las acciones se toman o son propuestas principalmente tras la ocurrencia de los accidentes, y *2*) las partes interesadas apelan al uso de la cooperación entre plataformas, en el entendido de que este concepto otorga cierta legitimidad a lo que se vaya a hacer para buscar una solución, dejando de lado la competitividad entre las plataformas. Por otro lado, el modelo de base de datos *hash*, que caracteriza al GIFCT, se empezó a usar primero para combatir el contenido relacionado al abuso sexual infantil (CSAM) en las plataformas. En este sentido, Google, en su blog, explica las diferentes tecnologías que emplea para detectar y eliminar ese contenido, además, menciona cómo la base de datos *hash* permite actuar sin necesidad de ver las imágenes y cuidando la privacidad de los usuarios.

Pero ¿es la cooperación entre plataformas la panacea para la moderación de contenido en línea?

En principio, vemos cómo la cooperación permite que las pequeñas plataformas tengan acceso a recursos y herramientas para moderar contenido, lo cual es importante considerando que se encuentran en neta desventaja por su tamaño y recursos limitados, y que existe una fuerte tendencia de los usuarios maliciosos para mover sus operaciones en ellas, donde, con probabilidad, su contenido no sería apartado con la misma rapidez y facilidad que en otras plataformas. Ese fue el caso de JustPaste.it, plataforma pequeña que al afiliarse al GIFCT, vio simplificada su labor de combatir el contenido extremista difundido en su espacio por el ISIS (Islamic State of Iraq and Syria).

En su informe de transparencia de 2022 (GIFCT, 2021), el GIFCT enumera las acciones impulsadas por las diferentes partes interesadas que lo integran, entre las que señala: la adopción de una interpretación común sobre contenido terrorista y extremista; las mentorías auspiciadas por el programa Tech Against –que dieron como resultado la publicación de reportes iniciales de transparencia–, así como los esfuerzos de los grupos de trabajo, los cuales fomentaron espacios de diálogo y publicaron investigaciones y reportes en temas como la expansión de la taxonomía, metodologías para evaluar los algoritmos de difusión de contenidos y el protocolo de respuesta a crisis.

No obstante, el sistema, tal como lo conocemos hoy, ha sido criticado por diversos motivos. Para comenzar, carece de transparencia, garantías procesales y rendición de cuentas: ¿qué contenido está presente en la base de datos *hash*?, ¿existe un registro de quiénes agregan qué? El GIFCT contempla informes anuales de transparencia –como el antes referido–, y en 2021 publicó un reporte en el que trata la disposición para expandir dicha base y su taxonomía. Sin embargo, estos documentos no contienen información relacionada con los cuestionamientos planteados, como tampoco existen instancias que supervisen y auditen esta cooperación, aun cuando su modelo de gobernanza prevé un director ejecutivo, un consejo ejecutivo y un comité asesor independiente. El consejo está compuesto por las cuatro compañías fundadoras mientras que el comité cuenta con actores de la sociedad civil, gobiernos, academia y organizaciones intergubernamentales, y tienen como principal objetivo "identificar y recomendar prioridades y áreas de interés clave para promover la misión y los pilares de trabajo del Foro y evaluar los progresos realizados en relación con estas recomendaciones y su misión" (Global Internet Forum to Counter Terrorism, n.d. a). Por consiguiente, varias organizaciones no gubernamentales han reclamado la necesidad de diversificar el conjunto de actores involucrados en este sistema y de reorganizar sus labores con un enfoque más orientado a los derechos humanos (Llansó, 2020).

La poca transparencia que hasta ahora ha caracterizado al modelo aquí analizado conlleva serios problemas asociados a la censura. Si bien es cierto que las plataformas digitales pertenecientes al GIFCT incluyen en sus reportes anuales la cantidad de solicitudes de remoción o de entrega de datos para investigaciones oficiales recibidas por parte de los gobiernos–por ejemplo, los informes de TikTok (2022) y Meta (n.d.) correspondientes a 2022–, no siempre añaden detalles sobre la tipología del contenido removido ni las razones que argumentan la medida. Además, aplicar el modelo de cooperación a otras categorías no tan específicas como el CSAM, crea una "zona gris" donde podrían intervenir criterios individuales y condicionados por las reglas de negocios adoptadas por cada plataforma, lo cual incrementaría la diversidad de contenido moderado mediante un sistema que, antes de seguir con su expansión e implementación, requiere un análisis profundo. Cabe añadir que

dicho análisis ayudaría a no perpetuar la poca legitimidad y credibilidad en los modelos de moderación de contenido en línea.

Entre las categorías más generales están las campañas de influencia exterior, las cuales son definidas como "campañas de un Estado coordinadas para influir en uno o más aspectos específicos de la política de otro Estado, a través de los medios de comunicación, incluidas las redes sociales, mediante la producción de contenidos diseñados para parecer autóctonos del Estado objetivo" (Martin et al., 2023). El incremento exponencial de esta clase de categorías en los últimos años ha llevado a un modelo en el que las plataformas y los gobiernos comparten datos e información para poder combatirlas (Douek, 2020). No obstante, esta cooperación informal y *ad hoc*, surgida para contrarrestar un tipo de contenido s*ui generis*, carece de legitimidad al no contar con un ente externo de supervisión. En ese mismo sentido, la aplicación de la base de datos de *hashes* a estas categorías no tan específicas podría provocar problemas, sobre todo por la falta de métricas y la necesidad de proveer modos de auditar la exactitud de los *hashes*.

Hay escepticismo respecto a la forma en que estas cooperaciones se desarrollan, ya que en el afán de "hacer algo" tras una crisis, se puede llegar a adoptar medidas que favorezcan exclusivamente a las grandes plataformas. Asimismo, resulta preocupante el alcance de estas medidas y su factibilidad para las plataformas más pequeñas, ya que hace falta solucionar los desafíos que las empresas menores enfrentan ante el requerimiento de detección y remoción del contenido señalado, en tiempos prudentes y dictados por ley, en su gran mayoría (Douek, 2020). Las compañías que deseen pertenecer al GIFCT deben, entonces, cumplir con una serie de disposiciones, como la de tener la habilidad para "recibir, revisar y actuar tanto sobre las denuncias de actividades ilegales o que infrinjan las condiciones del servicio como sobre las apelaciones de los usuarios". Si consideramos las fuertes inversiones económicas y de capital humano que las grandes plataformas realizan en la lucha contra el contenido ilícito, vemos la grave desventaja en la que se encuentran las más pequeñas, por lo que este organismo debe enfocarse en brindarles mayor asistencia, más allá de la base de datos de *hashes*.

Cabe destacar que estas desventajas fueron abordadas en 2021 por BSR en el GIFCT Human Rights Assessment (BSR, 2021). En esta ocasión BSR (Business for Social Responsibility)––organización que se especializa en temas de asesoría con enfoque en la sostenibilidad––evaluó el GIFCT basándose en los Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos (UNGP), concluyendo con la remisión de un informe que reúne unas 47 recomendaciones en torno a 9 temas, incluyendo transparencia, modelo de gobernanza y membresía, entre otros. Al respecto, en el tema de

gobernabilidad se valoró la inviabilidad del actual modelo de gobernanza del GIFCT a corto y mediano plazo, y se sugirió su transformación en una iniciativa basada en el modelo de múltiples partes interesadas, a un plazo de dos años, con el argumento de que el cambio favorecería la transparencia y el enfoque colaborativo, aspectos indispensables para los objetivos relacionados a los derechos humanos.

Al mismo tiempo, el informe enumera una serie de recomendaciones para mejorar los canales de comunicación entre el comité asesor y el consejo ejecutivo; fomentar la inclusión de las múltiples partes interesadas, considerando las minorías y grupos vulnerables; consensuar una definición de contenido de tipo terrorista y de violencia extremista, así como transparentar las actuaciones de los miembros mediante el uso de la base de datos de *hashes*. Sobre esta última, se incluyen sugerencias tales como la revisión y aprobación en forma manual antes de agregar nuevos *hashes,* los mecanismos de apelación y el permiso de consulta a otros actores, como la academia.

5. Repensar la cooperación entre plataformas: el *white paper* de Global Partners Digital

De acuerdo con lo analizado hasta este punto, existen varias iniciativas impulsadas por las diferentes partes interesadas para abordar la moderación de contenidos. La cooperación entre plataformas es un mecanismo que, con frecuencia, es apoyado por los sectores que buscan unificar criterios; la moderación de contenido en línea, al igual que otras áreas en la esfera digital, requiere un enfoque innovador y algo alejado de las prácticas conservadoras que se han estado aplicando hasta ahora. En este aspecto, vale la pena citar la propuesta hecha por el Global Partners Digital (GPD) en 2018, y reflejada en un documento técnico (*white paper*) dirigido a tres actores: gobierno, sociedad civil y plataformas digitales, por medio del cual se busca “proponer un modelo que respete los derechos derechos humanos y responda al legítimo interés de los gobiernos en los contenidos ilícitos y nocivos” (Bradley & Wingfield, 2018, p. 4).

En ese documento se esbozan las definiciones de contenido nocivo (*harmful content*) y contenido ilícito (*unlawful content*); este último es el que se encuentra prohibido en forma expresa por el derecho internacional de los derechos humanos, mientras que se entiende como nocivo aquel contenido pasible de restricción, según lo dictado por el artículo 19 del Pacto Internacional de Derechos Civiles y Políticos, sea por atentar contra la seguridad nacional y el orden público como contra los derechos o reputación de los demás. Además, la propuesta del GPD se centra en tres acciones: definición de términos de servicio (TdS),

implementación de los TdS y mecanismos de reclamación y reparación.

La razón de ser de los términos de servicio radica en el requerimiento de transparentar qué contenido es susceptible de restricción o remoción y en qué circunstancias. En el documento en cuestión se indica que los TdS deben ser precisos y de fácil acceso, además, deben categorizar las tipologías de contenido y detallar su interpretación, estar sujetos a revisión periódica y su desarrollo debe implicar "consulta y compromiso con una gama de partes interesadas relevantes" (Bradley & Wingfield, 2018, p. 14).

La implementación de los TdS prevé mecanismos sencillos para que los usuarios notifiquen el contenido que infrinja los parámetros y este sea sometido a un *triage* adecuado, elaborado por un equipo especializado, para determinar a qué categoría pertenece. El artículo referido hace énfasis en la necesidad de destinar recursos para que la determinación se realice en forma oportuna y sugiriendo medidas como capacitaciones, formalización de procesos para asegurar la calidad de la toma de decisiones y contratación de expertos externos. Asimismo, instruye tiempos prudentes para analizar las denuncias y contactar a los autores a fin de que, si aplicase, tengan oportunidad de justificar la existencia de ese contenido. La misma celeridad y pertinencia en los tiempos se aplicaría a la comunicación hacia las partes interesadas (denunciante y autor) respecto a la notificación recibida y la decisión tomada.

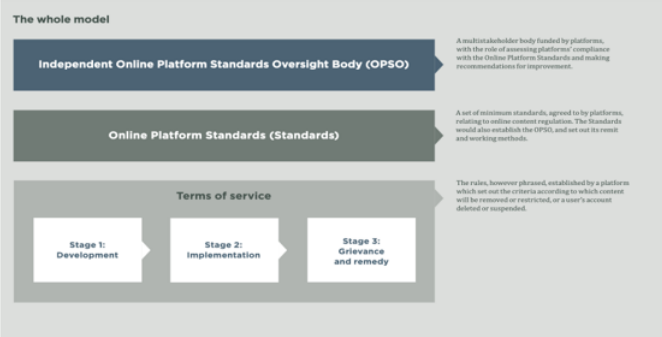
Las decisiones finales podrían ser impugnadas por medio de mecanismos de reclamación y reparación, que cumplan con el principio 31 de los Principios rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos, tratándose, así, de herramientas compatibles con los derechos humanos, accesibles, legítimas, transparentes, equitativas y que permitan un proceso de aprendizaje y mejora continuos, sin dejar de lado la posibilidad de optar por procesos judiciales externos a la esfera digital.

Otro planteamiento interesante del GPD es la creación de un organismo de supervisión. De manera precisa, se trataría de un organismo de alcance global, independiente, financiado por las plataformas y cuya membresía estaría determinada por un grupo de múltiples partes interesadas. Tal propuesta se debe, principalmente, a la disposición para garantizar la representación del interés público, la transparencia y la rendición de cuentas, las cuales se lograrían mediante la implementación del enfoque de múltiples partes interesadas y de una serie de estándares desarrollados de modo independiente. Cabe precisar que estos estándares estarían sujetos a revisiones periódicas y el mismo organismo velaría por su cumplimiento. En cuanto a la metodología que sería utilizada para la moderación de

contenidos, el documento menciona la revisión frecuente de los algoritmos y procesos automatizados, acompañada de supervisión humana constante.

Ante un panorama mundial tan impredecible donde la libertad de expresión a menudo se ve amenazada por medidas unilaterales que comprometen la navegación y el uso seguro de las plataformas (O'Sullivan, 2022), los puntos propuestos por GDP brindan un enfoque innovador y respetuoso de los derechos humanos.

Figura 2. Modelo sugerido por el GPD



Fuente: https://gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf

6. Hacia una cooperación más transparente y respetuosa de los derechos humanos

A partir del análisis realizado en esta investigación, procederemos a formular una propuesta para un modelo de cooperación con enfoque colaborativo y que goce de mecanismos que garanticen la transparencia, la legitimidad y la supervisión independiente, sin dejar atrás la perspectiva de derechos humanos en su concepción y desempeño.

En primer lugar, es necesario establecer un organismo de supervisión, el cual se encargaría de garantizar que las políticas y prácticas de moderación de las plataformas participantes cumplan con los estándares de derechos humanos, promoviendo, a la par, la transparencia y la rendición de cuentas de las plataformas. También, abogaría por la creación y adopción de estándares mínimos e indicadores aplicables a todas las plataformas involucradas, para

garantizar un marco común, siendo responsable de publicar informes periódicos sobre los avances y acciones realizadas, que incluyan estadísticas relevantes y los criterios utilizados en la toma de decisiones. Además, tendría la facultad de crear comisiones *ad hoc* para que revisen las decisiones de las plataformas; esta atribución sería similar a las facultades actuales del Consejo Asesor de Contenido de Meta.

Asimismo, el organismo en cuestión procuraría fomentar la participación de los diferentes actores en la revisión periódica de las políticas y prácticas, la publicación de informes de transparencia y la cooperación para el fortalecimiento continuo de la estructura, por lo que una de sus facultades sería la creación de espacios en los que sea factible compartir buenas prácticas entre las plataformas, ayudando, así, a las más pequeñas. Además, tendría la tarea de consensuar una interpretación común para el tipo de contenido ilegal que se pretende erradicar –como hizo el GIFCT con su taxonomía–, lo cual ayudaría a no replicar las dificultades que, por lo general, surgen cuando se aborda contenido no tan específico. También, favorecería las auditorías y evaluaciones independientes para garantizar la eficacia de la cooperación y el cumplimiento de los derechos humanos en sus políticas y procedimientos, y acompañaría a las plataformas en los procesos de adecuación de sus términos de referencia y políticas de contenido, según lo decidido en los espacios de discusión auspiciados.

El modelo de gobernanza a adoptar para el funcionamiento del organismo de supervisión propuesto se basaría en el principio de múltiples partes interesadas, el cual se caracteriza por "personas y organizaciones de diferentes ámbitos que participan unos junto a otros para compartir ideas o desarrollar políticas consensuadas" (Internet Society, 2016). Organizaciones del ecosistema de internet, como Internet Corporation for Assigned Names and Numbers (ICANN), lo han implementado en su día a día, y el mismo informe del BSR invita al GIFCT a considerar la transición de la junta operativa hacia este modelo, que finalmente ha sido incluido como meta a largo plazo en la hoja de ruta de dicho foro.

Por otro lado, se podría estudiar la posibilidad de continuar promoviendo el uso de herramientas como la base de datos de *hashes*; no obstante, es necesario definir de manera clara y colaborativa el tipo de contenido ilícito a ser removido a través de esta herramienta, e incluir mecanismos que permitan supervisar y auditar el contenido de la base, así como valorar la posibilidad de proceder a la revisión y aprobación humana antes de agregar contenido. Es también indispensable seguir analizando y mejorando los demás instrumentos para la moderación de contenido, enfrentando desventajas como la precarización del trabajo de supervisión humana y los sesgos de la inteligencia artificial. En todo caso, la metodología a utilizar deberá ser medible, auditable, garante de transparencia y escogida en función de

las complejidades que caracterizan el tipo de contenido ilícito seleccionado.

En cuanto a las características del GIFCT, los tres pilares que lo rigen (*prevent, respond, learn*) ameritan ser adoptados como base de un nuevo modelo de cooperación. De igual manera, la conformación de grupos de trabajo, el apoyo a la investigación y la creación de definiciones y principios pueden resultar beneficiosos para un correcto desarrollo de la cooperación como tal, así como para garantizar un proceso de aprendizaje y de mejora continuo. El Content Incident Protocol también es un recurso valioso, que podría ser tomado en cuenta.

7. Conclusión

La cooperación entre plataformas es un mecanismo cada vez más implementado. No obstante, las principales problemáticas detectadas en su funcionamiento están relacionadas con la necesidad de transparentar las acciones e incentivar la rendición de cuentas, fomentar la diversidad de actores participantes en sus instancias y adoptar un modelo de gobernanza que vele por una representación equitativa, tenga bien definidas las tipologías de contenido que serían objeto de la cooperación y garantice los mecanismos de denuncia, sometimiento de quejas y reconsideración.

La propuesta del GPD y el GIFCT como caso de estudio representan un buen punto de partida para plantear un modelo de cooperación que cumpla con lo antes mencionado, alcanzando parámetros más claros y definidos en cuanto a la moderación de contenidos en línea basada en la cooperación entre plataformas. Esto sería posible con un organismo de supervisión, cuya gobernanza adopte las características del modelo *multistakeholder.* La revisión del enfoque de moderación mixta –inteligencia artificial y supervisión humana– y del funcionamiento de la base de datos de *hashes* también es oportuna para combatir sesgos y mejorar ambos sistemas.

Ante los acontecimientos más recientes en los que se ha visto que la toma de decisiones unilaterales, a causa de la ausencia de una regulación precisa y consensuada, ha estado comprometiendo la navegación y el uso seguro de las plataformas, es esencial la cooperación entre las plataformas. Lo deseable es que tal cooperación esté basada en la adopción de estándares mínimos y en un enfoque multidisciplinario que refleje las realidades locales y globales, amparando, a la vez, los derechos humanos. Con ello, se podrán evitar prácticas desleales y que vayan en detrimento de los derechos de los usuarios,

para, en cambio, seguir promoviendo la apertura, la interoperabilidad y la neutralidad de la internet.

Referencias

Access Now. (s.f.). *26 recomendaciones sobre gobernanza de contenido: una guía para legisladores, reguladores y encargados de políticas empresariales.* https://www.accessnow.org/cms/assets/uploads/2020/03/Recomendaciones-Gov-Contenidos.pdf

Access Now. (2022, 14 de noviembre). *Argentina: Firman acuerdos sobre derechos ante moderación de contenidos.* https://www.accessnow.org/argentina-acuerdos-derechos-en-plataformas-ante-moderacion-contenidos/

Andreu, A. (2022, 27 de agosto). *"La IA no es 100% fiable": la moderación de contenidos en redes sociales aún es deficiente y las plataformas no dan con la tecla correcta para que algoritmos y humanos convivan en paz.* Business Insider. https://www.businessinsider.es/maquinas-vs-humanos-como-modera-contenido-redes-sociales-1109935

Association for Progressive Communications. (2014). *Frequently asked questions on internet intermediary liability.* www.apc.org/en/pubs/apc%E2%80%99s-frequently-asked-questions-internet-intermed

Belli, L., Zingales, N., Curzi de Mendonça, Y., De Gregorio, G., Ducato, R., Oliveira da Cruz, L., Goanta, C., Gojkovic, T., Khoo, C., Kulk, S., Leerssen, P., Neves Lorenzon, L., Marsden, C., Marique, E., Oghia, M., Radsch, C., Stylianou, K., Weber, R., Wiersma, C., … & Zingales, N. (2021). *Glossary of platform law and policy terms*. FGV Direito Rio.

Bradley, C., & Wingfield, R. (2018). *A Rights-Respecting Model of Online Content Regulation by Platforms.* Global Partners Digital. https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf

Business for Social Responsibility, (2021). *Human Rights Assessment: Global Internet Forum to Counter Terrorism.*

Carta de los Derechos Fundamentales de la Unión Europea, 7 de diciembre de 2000.

Christchurch Call (n.d.). *About*. https://www.christchurchcall.com/about/

Convención Americana sobre Derechos Humanos, 22 de noviembre de 1969.

Declaración Americana sobre los Derechos y Deberes del Hombre, 1948.

Declaración Universal de Derechos Humanos, 10 de diciembre de 1948.

Douek, E. (2020) *The Rise of Content Cartels.* Knight First Amendment Institute. https://s3.amazonaws.com/kfai-documents/documents/704838d2ec/3.23.2021_Douek_MW–To-Print-.pdf

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy, 10*(1), 33-50. https://doi.org/10.1386/jdmp.10.1.33_1

Global Internet Forum to Counter Terrorism. (n.d. a). *GIFCT Independent Advisory Committee: Interim Terms of Reference.* https://gifct.org/wp-content/uploads/2021/09/GIFCT-IAC-Terms-of-Reference.pdf

Global Internet Forum to Counter Terrorism. (n.d. b). *Preventing terrorists and violent extremists from exploiting digital platforms*. https://gifct.org/

Global Internet Forum to Counter Terrorism. (2021a, June 18). *Content Incident Protocol (CIP)* [Vídeo]. Vimeo. https://vimeo.com/564637363 ; https://gifct.org/content-incident-protocol/

Global Internet Forum to Counter Terrorism. (2021b, June 18). *What is Hash-sharing?* [Vídeo]. Vimeo. https://vimeo.com/564638166 ; https://gifct.org/hsdb/

Global Internet Forum to Counter Terrorism. (2021c, July). *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps.* https://gifct.org/wp-content/uploads/2021/07/GIFCT-TaxonomyReport-2021.pdf

Global Internet Forum to Counter Terrorism. (2021d, July). *Transparency Report.* https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf

Global Internet Forum to Counter Terrorism. (2022e). 2022 GIFCT *Transparency Report.* https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

Internet Corporation for Assigned Names and Numbers. (n.d.). *ICANN's Multistakeholder Model*. https://www.icann.org/community

Internet Society. (2016, 28 de abril). *Gobernanza de Internet – Por qué funciona el enfoque de múltiples partes interesadas*.

https://www.internetsociety.org/es/resources/doc/2016/gobernanza-de-internet-por-que-funciona-el-enfoque-de-multiples-partes-interesadas/

Jasper, S. (2022, October 28). How we detect, remove and report child sexual abuse material. *Google*. https://blog.google/technology/safety-security/how-we-detect-remove-and-report-child-sexual-abuse-material/

Ley de Servicios Digitales, de 16 de noviembre de 2022 (2022) (Unión Europea).

Llansó, E. (2020, 30 de julio). *Human Rights NGOs in Coalition Letter to GIFCT.* Center for Democracy & Technology. https://cdt.org/insights/human-rights-ngos-in-coalition-letter-to-gifct/

Martin, D. A., Shapiro, J. N. & Ilhardt, J. G. (2023). *Online Political Influence Efforts Dataset* (Version 4). https://esoc.princeton.edu/publications/trends-online-influence-efforts

Meta. (n.d.). *Case studies.* https://transparency.fb.com/data/government-data-requests/case-studies/

Netzdurchsetzunggesetz, NetzDG, 2017. https://germanlawarchive.iuscomp.org/?p=1245

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2022). *Por un internet confiable – Regular las Plataformas Digitales para la Información como Bien Común.* https://www.unesco.org/es/internet-conference

Pacto Internacional de Derechos Civiles y Políticos, 23 de marzo de 1976.

Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos, 2011.

Radsch, C. (2017, April 20). *Proposed German legislation threatens broad internet censorship*. Committee to Protect Journalists. https://cpj.org/2017/04/proposed-german-legislation-threatens-broad-intern/

Radsch, C. (2020, September 30). *GIFCT: Possibly the Most Important Acronym You've Never Heard Of*. Just Security. https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of

S.1693 – 115th Congress (2017-2018): Stop Enabling Sex Traffickers Act of 2017. (2018, enero 10). https://www.congress.gov/bill/115th-congress/senate-bill/1693

Santisteban Galarza, M. (2022). Garantías frente a la moderación de contenidos en la Propuesta de Reglamento Único de Servicios Digitales. *Revista CESCO de Derecho de Consumo*, (41), 159-179. https://doi.org/10.18239/RCDC_2022.41.3103

Silva, I. (2022, 10 de noviembre). *Las máquinas no son suficientes*. Derechos Digitales. https://www.derechosdigitales.org/19666/las-maquinas-no-son-suficientes/

Tech Against Terrorism. (n.d.). *Case study: Using the GIFCT hash-sharing database on small tech platforms.* https://www.counterextremism.com/sites/default/files/TAT%20–%20JustPaste.it%20GIFCT%20hash-sharing%20Case%20study.pdf

Text – H.R.1865 – 115th Congress (2017-2018): Allow States and Victims to Fight Online

Sex Trafficking Act of 2017. (2018, abril 11). https://www.congress.gov/bill/115th-congress/house-bill/1865/text

TikTok. (2022, November 29). *Government Removal Requests Report.* https://www.tiktok.com/transparency/en-au/government-removal-requests-2022-1/

The Manila Principles on Intermediary Liability [Background paper]. (2015). https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf

The Santa Clara Principles On Transparency and Accountability in Content Moderation. (n.d.). Santa Clara Principles 2.0. https://santaclaraprinciples.org/

United Nations. (2016). *Joint Declaration on Freedom of Expression and countering violent extremism*. https://www.ohchr.org/en/statements/2016/05/joint-declaration-freedom-expression-and-countering-violent-extremism?LangID=E&NewsID=19915

United Nations, Organization for Security and Co-operation in Europe, Organization of American States, & African Commission on Human and Peoples' Rights. (2017). *Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda*. https://www.osce.org/files/f/documents/6/8/302796.pdf

What is SESTA/FOSTA? (2020, May 8). Decriminalize Sex Work. https://decriminalizesex.work/advocacy/sesta-fosta/what-is-sesta-fosta/

Wingfield, R. (2018). *Riesgos y responsabilidades de la moderación de contenidos en las plataformas en línea.* Open Global Rights. https://www.openglobalrights.org/risks-and-responsibilities-of-content-moderation-in-online-platforms/?lang=Spanish

Biografía de la autora:

Federica Tortorella Casilla es Licenciada en Derecho por la Universidad Nacional Pedro Henríquez Ureña (UNPHU) con un Máster en Gerencia Integral de Riesgo por el Instituto Europeo de Posgrado. Actualmente se desempeña como Oficial de Políticas para LACTLD, la asociación que reúne a los ccTLD de Latinoamérica y el Caribe. Desde 2016 está involucrada en temas relacionados a la gobernanza de internet, participando en las cohortes de programas como el Youth@IGF (actual Youth Ambassador Program), ICANN Next Gen y la Diplomatura en Gobernanza de Internet (DiGI) de 2022. También ha realizado voluntariado en espacios relacionados a los derechos digitales.