



f

Manual de moderación de contenidos y libertad de expresión

Agosto 2023



ARTICLE 19

T: +44 20-7324 2500
F: +44 20-7490 0566
E: info@article19.org
W: www.article19.org
Tw: [@article19org](https://twitter.com/article19org)
Fb: facebook.com/article19org

© ARTICLE 19, 2023

Esta publicación se realizó con el apoyo financiero de la Unión Europea y la UNESCO. Las denominaciones empleadas y la presentación del material en esta publicación no implican ninguna opinión por parte de la UNESCO o de la Unión Europea sobre la condición jurídica de ningún país, territorio, ciudad o zona, o de sus autoridades, ni respecto de la delimitación de sus fronteras o límites.

Los autores son responsables de la elección y presentación de los hechos contenidos en la publicación y de las opiniones expresadas en esta, que no son necesariamente las de la Unión Europea o la UNESCO y no comprometen a las Organizaciones.

Esta obra se ofrece bajo licencia Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0. Se puede copiar, distribuir y mostrar libremente esta obra, así como realizar obras derivadas, siempre y cuando:

- 1) se dé crédito a ARTICLE 19;
- 2) no se utilice esta obra con fines comerciales
- 3) se distribuya cualquier obra derivada de esta publicación bajo una licencia idéntica a la presente.

Para acceder al texto legal completo de esta licencia, consultar:

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Contenido

Introducción	4
El poder de las empresas de redes sociales sobre la libertad de expresión	4
Objetivo y estructura de este manual	5
Sobre el proyecto	6
Estándares internacionales de derechos humanos aplicables	7
Panorama general	7
Garantías del derecho a la libertad de expresión	7
Limitaciones del derecho a la libertad de expresión	8
Libertad de expresión y 'discurso de odio'	10
<i>'Discurso de odio' que debería estar prohibido</i>	11
<i>'Discurso de odio' que puede estar prohibido</i>	12
<i>'Discurso de odio' legal</i>	12
Libertad de expresión y 'desinformación'	14
Responsabilidad de las empresas de redes sociales de respetar los derechos humanos	17
Moderación de contenidos en la práctica	23
Terminología clave	23
Términos de servicio y normas comunitarias	25
Preocupaciones que plantea la regulación del discurso por contrato	27
<i>Reducción de los estándares de libertad de expresión</i>	27
<i>Falta de transparencia y rendición de cuentas</i>	28
<i>Falta de garantías procesales y recursos de reposición</i>	30
<i>Elusión del Estado de Derecho</i>	31
El papel de los marcos regulatorios	32
<i>Enfoque tradicional en la regulación de la responsabilidad de intermediarios</i>	33
<i>Tendencia hacia una mayor regulación de las plataformas en línea</i>	35
<i>La regulación debe adoptar un enfoque basado en los derechos humanos e incluir los mercados digitales</i>	36
<i>Medios de comunicación y moderación de contenidos</i>	39
Procesos de moderación de contenidos	41
<i>Las deficiencias de la automatización</i>	42
<i>Falta de precisión y fiabilidad</i>	42
<i>Amplificación de sesgos</i>	44
<i>Falta de transparencia y rendición de cuentas</i>	45
<i>Denuncias de usuarios y 'marcadores de confianza'</i>	45
La necesidad de una comprensión más profunda del contexto	46
Conclusión	47
Bibliografía	49
Notas finales	50

Introducción

Esta publicación hace parte del proyecto **Social Media 4 Peace** de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), financiado por la Unión Europea (UE).

El poder de las empresas de redes sociales sobre la libertad de expresión

En sus inicios, las plataformas de redes sociales se consideraban una poderosa fuerza para el bien, que favorecía la libertad de expresión, facilitaba las conexiones entre las personas y lideraba una revolución democrática en todo el mundo. Esta percepción ha cambiado. Hoy en día, un número muy reducido de grandes empresas de redes sociales actúan como guardianes, controlando lo que un gran número de personas pueden ver o decir en Internet. Tienen un impacto directo en la dinámica de la distribución de contenidos, así como en la diversidad de medios en línea y en la libertad de expresión.

Este poder e influencia significativos se suman al hecho de que los modelos de negocio de las mayores empresas de redes sociales a menudo se basan en la recolección de grandes cantidades de datos sobre sus usuarios y sus hábitos en línea (datos de comportamiento) y en la monetización de estos datos a través de la publicidad en línea (dirigida). Esto interfiere significativamente con el derecho a la privacidad de los usuarios y puede tener un efecto negativo en la libertad de expresión. Es especialmente preocupante la propagación del 'discurso de odio' y la 'desinformación' en las plataformas en línea.¹ Las empresas de redes sociales han sido acusadas de priorizar el lucro sobre la seguridad de los usuarios mediante algoritmos que promueven el consumo de contenido perjudicial, incluyendo 'discurso de odio' y 'desinformación'. Cada vez se pide más a las empresas de redes sociales reforzar su labor de moderación de contenidos para enfrentar este tipo de contenidos problemáticos.

Objetivo y estructura de este manual

La moderación de contenidos incluye los distintos conjuntos de medidas y herramientas que las empresas de redes sociales utilizan para hacer frente a los contenidos de sus plataformas que son ilegales o que infringen las propias normas comunitarias de las empresas. Está influenciada por una serie de factores, que van desde la implementación de los modelos de negocio de las empresas de redes sociales hasta la presión de anunciantes y reguladores para evitar discursos socialmente indeseables, perjudiciales o ilegales, y la necesidad de proteger la libertad de expresión en línea.

Este manual, centrado en las mayores plataformas de redes sociales, ofrece un panorama conciso de la situación actual de la moderación de contenidos y de algunos de los principales problemas que surgen desde el punto de vista de la libertad de expresión.

Este manual está estructurado de la siguiente manera:

- En primer lugar, expone los estándares para la protección de la libertad de expresión en línea que se aplican a la moderación de contenidos, con un enfoque específico en el 'discurso de odio' y la 'desinformación'.
- En segundo lugar, aborda la relación contractual entre los usuarios y las principales empresas de redes sociales a través de los términos de servicio y las normas comunitarias que rigen el discurso en línea, así como los problemas que se presentan al regular el discurso por contrato.
- En tercer lugar, explica el papel de los marcos regulatorios en la moderación de contenidos, en particular el concepto de responsabilidad de intermediarios y la reciente tendencia hacia una mayor regulación de las empresas de redes sociales.
- Por último, analiza los procesos de moderación de contenidos típicamente aplicados por las principales empresas de redes sociales -incluidos los sistemas automatizados, los moderadores humanos y las denuncias de usuarios y terceros-, con especial atención a las deficiencias de los sistemas automatizados de moderación de contenidos.

Para un análisis de los desafíos específicos que rodean la desconexión entre las prácticas de moderación de contenidos de las principales empresas de redes sociales y las comunidades locales en las que se producen y distribuyen los contenidos moderados — basado en un estudio de las prácticas actuales en Bosnia y Herzegovina, Kenia e Indonesia—, ver el informe de ARTICLE 19 [*Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society*](#) [*Moderación de contenidos y libertad de expresión: cerrando brechas entre las redes sociales y la sociedad civil local.*].

Sobre el proyecto

Este manual hace parte del proyecto **Social Media 4 Peace** que la UNESCO y ARTICLE 19 están llevando a cabo en Bosnia y Herzegovina, Kenia, Indonesia y Colombia con el apoyo de la UE. El objetivo general del proyecto es reforzar la resiliencia de las sociedades frente a los contenidos potencialmente perjudiciales difundidos en línea, en particular el 'discurso de odio' y la 'desinformación', al tiempo que se protege la libertad de expresión y se contribuye a la promoción de narrativas de paz a través de las tecnologías digitales, especialmente las redes sociales. La contribución de ARTICLE 19 al proyecto se centra en las preocupaciones planteadas por las prácticas actuales de moderación de contenidos en las principales plataformas de redes sociales de los cuatro países objetivo.

Estándares internacionales de derechos humanos aplicables

Panorama general

¿Cómo se aplican los estándares internacionales de derechos humanos y libertad de expresión a la moderación de contenido? La moderación de contenidos está influenciada y determinada principalmente por las acciones de las empresas de redes sociales, por un lado, y por los actores estatales, en particular a través de leyes y regulaciones, por el otro. Ambos tienen responsabilidades en virtud del derecho internacional de los derechos humanos, incluida la protección de la libertad de expresión, aunque en distintos grados.

En primer lugar, esta sección delineará los estándares aplicables para la protección de la libertad de expresión en línea que deben guiar cualquier medida adoptada por los Estados y las empresas de redes sociales en la moderación de contenidos. En segundo lugar, ofrecerá brevemente una visión general de cómo se aplican los estándares internacionales de libertad de expresión al 'discurso de odio' y a la 'desinformación' —dos categorías de discurso que abarcan una amplia gama de expresiones, pero que carecen de una definición uniforme bajo el derecho internacional de los derechos humanos. En tercer lugar, explicará el alcance de las responsabilidades de las empresas de redes sociales en materia de derechos humanos, cómo difieren de las de los Estados y qué significa esto en la práctica para la moderación de contenidos.

Garantías del derecho a la libertad de expresión

El derecho a la libertad de expresión está protegido por el artículo 19 de la Declaración Universal de Derechos Humanos (DUDH),² y adquiere fuerza legal a través del Artículo 19 del Pacto Internacional de Derechos Civiles y Políticos (PIDCP)³ y en los tratados regionales.⁴

El alcance del derecho a la libertad de expresión es amplio. Exige que los Estados garanticen a todas las personas la libertad de buscar, recibir o difundir informaciones o

ideas de cualquier tipo, sin consideración de fronteras, por el medio de expresión de su elección. Los Estados tienen la obligación de no interferir en la circulación de información e ideas ni restringir indebidamente la expresión. Los Estados también tienen la obligación positiva de promover condiciones que favorezcan la libertad de expresión y de proteger a las personas de injerencias desproporcionadas por parte de entidades privadas.⁵ En este contexto, se ha argumentado que los Estados están obligados a adoptar medidas positivas para garantizar que el derecho a la libertad de expresión pueda disfrutarse en línea efectivamente, por ejemplo, introduciendo garantías procesales en el marco jurídico relativo a la retirada de contenidos en línea.⁶

En 2011, el Comité de Derechos Humanos de la ONU, órgano de vigilancia del cumplimiento por parte de los Estados del PIDCP, aclaró que el derecho a la libertad de expresión se aplica también a todas las formas de expresión electrónica y por Internet.⁷

La [Caja mundial de herramientas para los operadores judiciales: normas jurídicas internacionales sobre la libertad de expresión, el acceso a la información y la seguridad de los periodistas](#) de la UNESCO ofrece un módulo específico sobre los nuevos retos de la protección de la libertad de expresión en Internet.

Limitaciones del derecho a la libertad de expresión

Bajo los estándares internacionales de derechos humanos, los Estados pueden, excepcionalmente, limitar el derecho a la libertad de expresión, siempre que dichas limitaciones cumplan con los estrictos requisitos del test tripartito del Artículo 19(3) del PIDCP.⁸ Esto requiere que las limitaciones sean:

- **Prescritas por ley**
 - Cualquier restricción debe estar formulada con suficiente precisión para permitir que las personas regulen su conducta consecuentemente. No se permiten las restricciones excesivamente amplias.

- Por ejemplo, una norma que penalice la 'difusión de rumores de forma que pueda afectar al bienestar general del público' estaría abierta a muchas interpretaciones diferentes -incluidos los significados de 'rumor', 'difusión' o 'bienestar general'- y no cumpliría el estándar de calidad exigido por el test tripartito.

- **En búsqueda de un objetivo legítimo**

- Las restricciones solo están permitidas para (a) el respeto de los derechos o la reputación de los demás, y (b) la protección de la seguridad nacional, del orden público, de la salud pública o de la moral pública.
- Por ejemplo, las autoridades no están autorizadas para imponer restricciones al ejercicio del derecho a la libertad de expresión con el fin de garantizar el respeto por un 'texto religioso reconocido' o para proteger las religiones de la burla.

- **Necesarias y proporcionadas en una sociedad democrática**

- Las restricciones deben demostrar una conexión directa e inmediata entre la expresión y el interés protegido. Además, si una medida menos restrictiva es capaz de lograr el mismo propósito que una más restrictiva, debe aplicarse la medida menos restrictiva.
- Por ejemplo, imponer una pena de prisión por difamación constituiría una restricción desproporcionada de la libertad de expresión. La reparación por difamación debería ser proporcionada, en cambio, a través de acciones en el ámbito del derecho civil o de medidas alternativas, incluyendo las disculpas, correcciones y el uso del derecho de réplica. La sanción de prisión es demasiado severa para una conducta que implique un daño a la reputación de una persona. En general, la pena de prisión debería reservarse para los peores delitos de expresión, como la incitación al genocidio.

El test tripartito puede aplicarse a todas las medidas adoptadas por el Estado, incluidas las medidas legislativas, las políticas públicas o las sentencias dictadas contra particulares. En

cuanto a la moderación de contenidos, el test tripartito se aplicaría normalmente para evaluar los marcos normativos que rigen a las empresas de redes sociales o las solicitudes formuladas por los Estados para acceder a los datos de los usuarios o restringir contenidos.

Como se explicará con más detalle a continuación, las empresas de redes sociales también tienen responsabilidades de respetar los derechos humanos, incluyendo la libertad de expresión, y deben asegurarse de que sus productos y servicios sean acordes a los estándares internacionales de derechos humanos. Por lo tanto, el test tripartito también podría ser utilizado para analizar si los términos de servicio de las empresas o las decisiones individuales de moderación de contenido están alineados con los estándares internacionales de libertad de expresión.

Libertad de expresión y 'discurso de odio'

Abordar en detalle los diferentes tipos de 'discurso de odio' y cómo deben ser tratados dentro de un marco de derechos humanos va más allá del alcance de este manual.⁹ Por lo tanto, esta sección se limita a esbozar una serie de distinciones y principios básicos cuando se trata de restricciones permisibles del 'discurso de odio' según los estándares internacionales de libertad de expresión.

No existe una definición unánime de 'discurso de odio' en el derecho internacional de los derechos humanos. En términos sencillos, el 'discurso de odio' es cualquier expresión de odio discriminatorio hacia las personas. Y es la protección de la igualdad y el principio de no discriminación, también protegidos en el PIDCP, lo que motiva la mayoría de las respuestas contra el 'discurso de odio'. El principio de no discriminación protege a las personas de cualquier distinción, exclusión o restricción basada en una característica protegida. Algunas de ellas se enumeran en el artículo 26 del PIDCP; pueden ser la raza, el color, el sexo, el idioma, la religión, las opiniones políticas o de cualquier otra índole, el origen nacional o social, la posición económica, el nacimiento o cualquier otro estatus.

Sin embargo, el 'discurso de odio' definido en sentido amplio como cualquier expresión de odio discriminatorio hacia las personas no implica necesariamente una consecuencia

concreta. Esta definición del mínimo común denominador abarca una gama muy amplia de expresiones, incluidas las expresiones lícitas. Por lo tanto, esta definición es demasiado vaga para su uso en la identificación de expresiones que pueden ser legítimamente restringidas bajo el derecho internacional de los derechos humanos.

ARTICLE 19 propone dividir el 'discurso de odio' en tres categorías, de la siguiente manera:

'Discurso de odio' que debería estar prohibido

El derecho penal internacional y el Artículo 20(2) del PIDCP exigen que los Estados prohíban determinadas formas graves de 'discurso de odio', incluso con medidas penales, civiles y administrativas.

El artículo 20.2 del PIDCP exige que los Estados prohíban por ley "toda apología del odio nacional, racial o religioso que constituya incitación a la discriminación, la hostilidad o la violencia" (incitación).

El Plan de Acción de Rabat, que proporciona orientación autorizada a los Estados sobre el cumplimiento de sus obligaciones en virtud del artículo 20.2 del PIDCP,¹⁰ establece una prueba de umbral de seis partes para las expresiones consideradas delitos penales bajo el artículo 20 del PIDCP. Requiere tener en cuenta (1) el contexto social y político, (2) la condición del orador, (3) la intención de incitar a la audiencia contra un grupo objetivo, (4) el contenido y la forma del discurso, (5) el alcance de su difusión, y (6) la probabilidad de daño, incluida su inminencia.

Caso hipotético

En la antesala de unas elecciones presidenciales muy reñidas, el presidente en ejercicio pronuncia una serie de discursos ante grandes multitudes. Durante estos mítines, difunde el rumor de que los partidarios de la oposición, en su mayoría pertenecientes a otro grupo étnico, se están armando y representan una amenaza existencial para sus seguidores. A medida que aumenta la tensión, utiliza un lenguaje racializado, que evoca las consignas utilizadas en las matanzas masivas perpetradas en el país unas décadas

antes, instando a sus partidarios a tomar medidas urgentes para asegurar una victoria electoral.

En este caso, el presidente ha proferido un 'discurso de odio' que podría alcanzar el umbral de apología del odio que constituye incitación a la violencia. El presidente entiende y está explotando las tensiones étnicas de la sociedad y, como político influyente, sabe que el uso que hace de un término concreto sería entendido y probablemente empleado violentamente por personas de la multitud contra miembros del grupo étnico asociado con la oposición.

'Discurso de odio' que puede estar prohibido

Los Estados pueden prohibir otras formas de 'discurso de odio', siempre que cumplan con los requisitos del Artículo 19(3) del PIDCP. Entre ellas se incluirían las manifestaciones individualmente dirigidas de amenazas, agresiones o acoso motivadas por prejuicios.

Caso hipotético

Una pareja del mismo sexo, ambas mujeres, es abordada en un tren por otro pasajero que comienza a gritarles insultos sexistas y homofóbicos, causándoles un temor razonable de violencia física.

En muchas jurisdicciones, este incidente sería, apropiadamente, perseguido como un delito motivado por prejuicios. El acto expresivo del pasajero agresor entra dentro de nuestra amplia tipología de 'discurso de odio' y también constituye un crimen de agresión. La amenaza plausible de violencia en la expresión la convierte en una conducta delictiva y, dado que está caracterizada por prejuicios, el contenido de la expresión es también una prueba de la motivación por prejuicios.

'Discurso de odio' legal

Debería estar protegido de restricciones en virtud del Artículo 19(2) del PIDCP, pero aun así plantea preocupaciones en términos de intolerancia y discriminación y merece una respuesta crítica por parte del Estado.

Caso hipotético

Un adolescente con un pequeño número de seguidores en Twitter tuitea un chiste ofensivo y sexista que trivializa la desaparición y probable asesinato de una estudiante local. Se provoca una fuerte reacción crítica contra el chico en Internet, que termina borrando el tuit. Aunque el comentario sea ofensivo y refleja un problema más amplio de misoginia en la sociedad, no pretendía incitar a ninguna conducta perjudicial contra un grupo concreto y, en cualquier caso, no tiene ese tipo de influencia sobre sus seguidores. Este tipo de 'discurso de odio' puede justificar la intervención moderada por parte de agentes locales en posiciones de autoridad, como los profesores de la escuela del chico u otros líderes de la comunidad, pero no justifica que el Estado imponga sanciones u otras restricciones.

Es importante tener en cuenta estas distintas categorías de 'discurso de odio' a la hora de evaluar las restricciones a la libertad de expresión, en particular las impuestas por los Estados, incluidas las relacionadas con cuestiones de moderación de contenidos.

Como se explica más detalladamente a continuación, las empresas de redes sociales suelen retirar contenido protegido bajo los estándares internacionales de libertad de expresión. La eliminación a menudo ocurre mediante la aplicación de las políticas de las plataformas sobre 'discurso de odio' y otras relacionadas. TikTok, por ejemplo, no permite 'ninguna ideología que incite al odio', incluyendo la 'misoginia'. Una prohibición absoluta de la misoginia no cumpliría con los requisitos de los artículos 19 y 20 del PIDCP. [TikTok tampoco permite](#) 'negar acontecimientos históricos debidamente documentados que perjudicaron a grupos basados en un atributo protegido' y afirma que ofrece 'algunas protecciones relacionadas con la edad'.¹¹

Las políticas pueden variar considerablemente de una plataforma a otra. Por ejemplo, mientras que [TikTok](#) incluye el 'género' y la 'identidad de género' en su lista de características protegidas, [Meta](#) sólo incluye la 'identidad de género'.

Si bien las empresas de redes sociales tienen derecho como entidades privadas a adoptar normas comunitarias más estrictas que los requisitos mencionados anteriormente, sus

términos de servicio y decisiones de moderación de contenidos deben, al menos, estar en consonancia con las normas y principios internacionales de derechos humanos.

Ejemplo

En el caso del 'poema ruso', el Consejo Asesor de Contenidos de Meta -un mecanismo establecido por Meta para revisar las decisiones de moderación de contenidos en casos seleccionados y proporcionar orientación sobre las políticas generales de moderación de contenidos de Meta a través de opiniones consultivas sobre políticas- evaluó la decisión de Meta de eliminar una publicación de Facebook de abril de 2022. La publicación fue realizada después de la invasión ilegal de Rusia a Ucrania y comparaba al ejército ruso en Ucrania con los nazis, citando un poema que llama al asesinato de fascistas. El Consejo afirmó que '[p]ara evaluar los riesgos planteados por los contenidos violentos o que incitan al odio, el Consejo [Asesor] generalmente se guía por la prueba de seis partes descrita en el [Plan de Acción de Rabat](#), que aborda la apología al odio nacional, racial o religioso que constituye incitación a la hostilidad, la discriminación o la violencia'.

En este caso, el Consejo Asesor consideró que, a pesar del contexto de conflicto armado en curso entre Rusia y Ucrania y de las referencias culturales intensas empleadas por el usuario, era poco probable que la publicación -una advertencia contra un ciclo de violencia- causara un daño. [El Oversight Board concluyó](#) que la eliminación inicial de contenidos no era necesaria.

Libertad de expresión y 'desinformación'

Al igual que el 'discurso de odio', los conceptos de 'desinformación', 'información errónea', propaganda e 'información falsa' carecen de una definición acordada en el derecho internacional o regional de los derechos humanos.¹² En general, los intentos de definir estos conceptos en legislaciones nacionales y estándares regionales se centran en prohibiciones de información 'falsa' o 'engañosa' que pueda causar cierto 'daño' o perjuicio.¹³

Dado que las restricciones a la 'desinformación' implican limitaciones a la libertad de expresión, cualquier respuesta jurídica y política a estos problemas debe basarse en estándares internacionales de libertad de expresión. En particular, debe cumplir con el test tripartito del Artículo 19(3) del PIDCP. Las restricciones a la 'desinformación' sólo se permiten si están explícitamente vinculadas a la protección de los fines legítimos

especificados en el Artículo 19(3) –como el respeto de los derechos o la reputación de los demás y la protección de la seguridad nacional, el orden público o la salud o la moral públicas– o el Artículo 20 del PIDCP – que prohíbe la propaganda en favor de la guerra y toda apología del odio nacional, racial o religioso que constituya incitación a la discriminación, la hostilidad o la violencia. La mera falsedad o naturaleza engañosa de cierta información no basta para restringir su difusión. Además, para que las restricciones a la 'desinformación' sean compatibles con los estándares internacionales sobre libertad de expresión, deben aplicarse de manera proporcionada.

En general, las consecuencias perjudiciales de la 'desinformación' deben abordarse con medidas habilitadoras, como garantizar un entorno mediático libre, independiente y diverso y promover la exposición de las personas a la mayor diversidad posible de información. Por otra parte, las prohibiciones y otras restricciones legales a la difusión de información falsa se pueden prestar a abusos y tener un impacto devastador en el discurso político. Promulgar un deber legal de 'verdad' crea un poderoso instrumento para controlar los flujos de información e ideas, lo que puede ser una herramienta peligrosa en manos de las autoridades. Y, de hecho, los conceptos de 'desinformación', 'información errónea', propaganda e 'información falsa' han sido utilizados y abusados por quienes detentan el poder como medio de represión de la disidencia y de desacreditación de la información que no les gusta.

Aunque la legislación que criminaliza de alguna forma la difusión de 'noticias falsas' no es nada nuevo, la actividad regulatoria ha aumentado considerablemente en los últimos años. Gobiernos de todo el mundo han creado o actualizado delitos de 'desinformación', incluso mediante legislación supuestamente destinada a combatir los delitos cibernéticos, adoptada en el contexto de la pandemia de Covid-19 o dirigida a reprimir el cubrimiento de conflictos armados.¹⁴ Esta legislación y otras similares han sido utilizadas para detener y procesar a blogueros, periodistas y críticos de los gobiernos.

Ejemplo

En noviembre de 2022, el periodista senegalés Pape Alé Niang fue arrestado y detenido por las autoridades senegalesas. Se le acusó, entre otros cargos, de 'difusión de noticias falsas susceptibles de desacreditar a las instituciones públicas'. Los cargos se surgieron de su cobertura a las acusaciones de violación contra el principal líder de la oposición senegalesa, Ousmane Sonko, que habían creado tensiones políticas en el país. Pape Alé Niang fue finalmente liberado el 10 de enero de 2023, pero fue puesto bajo supervisión judicial. Pape Alé Niang es uno de varios periodistas que informan sobre asuntos de interés público que han sido procesados recientemente conforme a leyes que penalizan la difusión de 'información falsa'.

Ejemplo

En diciembre de 2022, una de las figuras más destacadas de la oposición rusa, Ilya Yashin, fue encarcelado durante ocho años y medio por condenar el asesinato de cientos de civiles ucranianos por parte de las fuerzas de ocupación rusas en Bucha en su canal de YouTube. En este video compartió imágenes y relatos de la escena tomados de la BBC y otros medios. El tribunal consideró que había difundido conscientemente información falsa sobre las fuerzas armadas rusas.

La condena se basó en una ley rusa promulgada tras la invasión ilegal de Ucrania que prevé penas de prisión de hasta 15 años para quienes sean declarados culpables de difundir 'noticias falsas' o cualquier información que las autoridades rusas consideren falsa en la cobertura relacionada con la guerra.

No solo resulta problemático que las autoridades se conviertan en árbitros de la verdad, sino también que las entidades privadas que controlan espacios de expresión cuasi públicos decidan cuál es la información 'correcta' y cuál la 'falsa' y supriman y restrinjan la información que consideren incorrecta. Por lo tanto, la regulación estatal que requiere que las plataformas tomen medidas para suprimir la información falsa o los términos de servicio que contienen prohibiciones generales de 'desinformación' constituirían interferencias indebidas a los derechos de libertad de expresión de los usuarios de la plataforma. Por lo tanto, una regulación estatal que obligue a las plataformas a tomar medidas para suprimir la información falsa o unos términos de servicio que contengan

prohibiciones generales de 'desinformación' constituirían injerencias indebidas en el derecho a la libertad de expresión de los usuarios de las plataformas.¹⁵

Responsabilidad de las empresas de redes sociales de respetar los derechos humanos

Los Estados son los principales titulares de obligaciones en virtud del derecho internacional de los derechos humanos. Tienen la obligación de respetar, proteger y cumplir los derechos humanos. Esto incluye el deber de proteger a las personas contra las violaciones de los derechos humanos cometidas por todos los agentes de la sociedad, incluidas las empresas. De hecho, los Estados deben prevenir, investigar, sancionar y reparar las violaciones de derechos humanos cometidas por agentes privados. Aunque algunas de las principales empresas de redes sociales -al igual que otras grandes corporaciones- tienen posiblemente más poder que algunos Estados y pueden afectar profundamente los derechos humanos de las personas y comunidades en donde operan, como empresas no están sujetas a las mismas obligaciones en materia de derechos humanos que los Estados.

Sin embargo, los Principios Rectores sobre Empresas y Derechos Humanos: Puesta en Práctica del Marco de las Naciones Unidas 'Proteger, Respetar y Remediar' (los Principios Rectores) reconocen que las empresas tienen la responsabilidad de respetar los derechos humanos, independientemente de las obligaciones estatales o de su implementación.¹⁶

En particular, los Principios Rectores recomiendan a las empresas:¹⁷

- hacer una declaración pública de su compromiso de respetar los derechos humanos, respaldada por la alta dirección o la dirección ejecutiva;
- implementar procesos de debida diligencia y realizar [evaluaciones de impacto sobre los derechos humanos](#) para identificar, prevenir y mitigar cualquier posible impacto negativo sobre los derechos humanos derivado de sus operaciones;

- incorporar salvaguardias para los derechos humanos desde el diseño para mitigar los impactos adversos y actuar colectivamente para reforzar su influencia frente a las autoridades gubernamentales;
- hacer seguimientos y comunicar sobre los rendimientos, los riesgos y las exigencias de los gobiernos;
- y ofrecer acceso a soluciones cuando se produzcan impactos adversos sobre los derechos humanos.

Más concretamente, las empresas de redes sociales pueden tener que evaluar y mitigar los efectos potencialmente negativos sobre los derechos humanos derivados de su respuesta a las solicitudes gubernamentales de retirada o acceso a datos o de sus prácticas de recolección de datos, curación de contenidos o publicidad dirigida, entre otros.

En lo que respecta a las prácticas de moderación de contenidos, las empresas de redes sociales deben, entre otras cosas, establecer términos de servicio claros e inequívocos acordes a las normas y principios del derecho internacional de los derechos humanos¹⁸, producir reportes de transparencia acerca de las solicitudes realizadas por los gobiernos¹⁹, garantizar que las sanciones por el incumplimiento de los términos de servicio sean proporcionales y ofrecer recursos efectivos a los usuarios afectados por vulneraciones a los términos de servicio²⁰.

La sociedad civil ha hecho recomendaciones específicas sobre la responsabilidad de respetar los estándares internacionales de derechos humanos que tienen las empresas de redes sociales. Por ejemplo, los [Principios de Manila sobre Responsabilidad de los Intermediarios](#) establecen que las prácticas de restricción de contenido de las empresas deben satisfacer el *test* de necesidad y proporcionalidad bajo las normas de derechos humanos (Principio 4) y deben ofrecer a los usuarios mecanismos de denuncia para apelar las decisiones de las empresas (Principio 5(c)).

Asimismo, algunas empresas de redes sociales se han comprometido públicamente a guiarse por ciertos principios de derechos humanos, ya sea a través de la adopción de una política de derechos humanos, como Meta, o de su pertenencia a iniciativas multipartitas como la [Global Network Initiative](#), cuyos miembros se comprometen a colaborar en avanzar los derechos de los usuarios a la libertad de expresión y privacidad.

Para comprender hasta qué punto la responsabilidad de respetar las normas del derecho internacional de los derechos humanos puede guiar las decisiones individuales de moderación de contenido, puede ser instructivo leer algunas decisiones del Consejo Asesor de Contenidos de Meta pues sus fuentes de autoridad son, tanto la política de contenidos de Meta, como sus responsabilidades de acuerdo con los Principios Rectores de las Naciones Unidas.

Ejemplo

El Consejo Asesor de Contenidos evaluó un caso relativo a la retirada de contenido publicado por un usuario de Instagram en enero de 2021 en el que aparecía una imagen de Abdullah Öcalan con un comentario en el que se decía que era hora de hablar de ponerle fin al aislamiento al que se encuentra sujeto Öcalan en prisión. Öcalan es uno de los miembros fundadores del Partido de los Trabajadores del Kurdistan (PKK), considerado una organización terrorista en Turquía, y se encuentra en prisión desde 1999. Tanto el PKK como Öcalan son calificados como entidades peligrosas según la política de Facebook sobre Personas y organizaciones peligrosas.

ARTICLE 19 presentó comentarios públicos sobre este caso defendiendo que cualquier restricción a la libertad de expresión en Facebook debe estar guiada por los estándares de derechos humanos, que las políticas que restringen el contenido sobre la base de que puede incitar a la actividad terrorista deben operar en el marco de una terminología estrictamente definida y que existe una discrepancia significativa entre el enfoque de derechos humanos sobre el 'contenido terrorista' y las normas comunitarias de Facebook, que se centran excesivamente en oradores u organizaciones.

El Consejo Asesor de Contenidos encontró que la remoción del contenido fue incompatible con el compromiso de Meta de respetar los derechos humanos en tanto la discusión sobre las condiciones de cualquier individuo detenido constituye contenido protegido; las normas comunitarias no eran lo suficientemente claras para cumplir con el test de legalidad y la remoción del contenido era innecesaria y desproporcionada dado que en el caso no se indicaba ningún apoyo a los actos violentos cometidos por Öcalan o el PKK.

La jurisprudencia sobre el rol que tienen los derechos humanos en las decisiones de moderación de contenidos parece limitada. Una notable excepción es la sentencia dictada por el Tribunal de Roma en *Facebook v. CasaPound*.

Ejemplo

El Tribunal de Roma desestimó el recurso de Meta (entonces Facebook) contra una medida cautelar que le ordenaba reactivar la cuenta del partido de extrema derecha italiano CasaPound. Meta había desactivado la cuenta sin previo aviso ni explicación y argumentó ante el Tribunal que sus acciones eran legítimas porque la cuenta incluía contenidos que constituían 'discursos de odio' e incitación a la violencia, en violación de los términos de uso de Meta.

El Tribunal falló en favor de CasaPound y sostuvo que el contrato celebrado con Meta, pese a ser un contrato de Derecho civil común, debía interpretarse de conformidad con la Constitución italiana, incluido el derecho a la libertad de expresión, ya que Meta desempeñaba de facto una función sistémica a efectos de participación política.

El Tribunal rechazó el argumento de Meta de que la suspensión 'sancionaba' el hecho de que CasaPound fuera una organización política intrínsecamente contraria a la Constitución y a la normativa de derechos humanos. Sostuvo que no correspondía a Meta determinar si CasaPound era un actor legítimo, teniendo en cuenta además que el partido CasaPound no había sido declarado ilegal por las autoridades italianas competentes. El Tribunal consideró además que la relación contractual en este caso se había extinguido indebidamente y que los contenidos compartidos por CasaPound no alcanzaban un grado de gravedad tal que justificara su prohibición absoluta.²¹

En un caso similar, el Tribunal Federal Supremo de Alemania encontró en 2021 que Meta (entonces Facebook) estaba sujeta a los derechos fundamentales reconocidos en Alemania, incluyendo el derecho a la libertad de expresión.

Ejemplo

El Tribunal Supremo Federal de Alemania determinó, en un caso relacionado con la eliminación de contenido y la suspensión de cuentas por Facebook debido a alegaciones de ‘discurso de odio’, que la compañía debía encontrar un balance en sus términos de servicio entre los derechos fundamentales en conflicto: el derecho a la libertad de expresión y el derecho de Meta a ejercer una profesión libremente, ambos garantizados en la Ley Fundamental alemana. Ello exigía que Meta informara a sus usuarios del retiro de sus publicaciones luego de hacerlo y avisara con anterioridad sobre el bloqueo de las cuentas, motivara estas acciones y permitiera que los usuarios respondieran, tras lo cual se debería adoptar una nueva decisión.

Los términos y condiciones de Meta para la remoción de contenido y bloqueo de cuentas no cumplía con estos requerimientos. En consecuencia, el Tribunal encontró que la empresa no tenía derecho a eliminar las publicaciones de los demandantes ni a bloquear sus cuentas de usuario.

Las empresas de redes sociales se enfrentan a menudo con exigencias legales de los gobiernos que no se encuentran en línea con los estándares internacionales de derechos humanos – por ejemplo, bloquear ciertas cuentas, retirar contenido o facilitar el acceso a datos de los usuarios. Estas demandas pueden basarse en cibercrimen, ‘desinformación’ u otras leyes que no cumplen los estándares internacionales sobre la libertad de expresión.

Como se explicó anteriormente, la responsabilidad de las empresas de respetar los derechos humanos es independiente de la voluntad de un gobierno de cumplir sus obligaciones de derechos humanos. El ex Relator Especial para la Libertad de Expresión de la ONU ha especificado que la responsabilidad de las empresas de respetar la libertad de expresión debe, como mínimo, incluir el deber de ‘participar en estrategias de prevención y mitigación que respeten en la mayor medida posible los principios de derechos humanos internacionalmente reconocidos cuando se encuentren con requisitos incompatibles establecidos en la legislación local’²². Las exigencias legales deben interpretarse e

implementarse de la forma más restringida posible para garantizar la menor restricción posible del derecho a la libertad de expresión²³. El Relator Especial sostuvo además que cuando las empresas reciben este tipo de solicitudes deben ‘solicitar aclaraciones o modificaciones; pedir el apoyo de la sociedad civil, otras empresas afines, autoridades gubernamentales competentes, órganos regionales o internacionales y otros interesados; y explorar todas las opciones que la ley ofrezca para impugnarlas’²⁴. Por último, las empresas deben ser transparentes sobre las solicitudes de los gobiernos y ofrecer detalles sobre el tipo de contenido objeto de las solicitudes (por ejemplo, difamación, ‘discurso de odio’, contenido relacionado con el terrorismo) y las medidas tomadas por las empresas (por ejemplo, remoción total o parcial, remoción global o específica a un país, suspensión de la cuenta, remoción concedida bajo los términos de servicio)²⁵.

Moderación de contenidos en la práctica

Terminología clave

Es útil referirse brevemente a cómo la moderación de contenido difiere de la curación de contenidos. Los sistemas de moderación de contenidos y de curación de contenidos pueden estar estrechamente relacionados, pero plantean cuestiones diferentes y usualmente se tratan por separado por los tomadores de decisión y reguladores.

Para los propósitos de este manual, nos basamos en las siguientes definiciones:

- **Moderación de contenido** incluye las distintas medidas y herramientas que una plataforma de redes sociales usa para hacer frente a los contenidos ilícitos y para aplicar sus normas comunitarias a los contenidos generados por los usuarios. Esto generalmente implica la denuncia por parte de los usuarios, ‘trusted flaggers’, o ‘filtros’; la remoción, el etiquetado, la desclasificación o desmonetización del contenido; o la desactivación de ciertas funciones.
- **Curación de contenidos** se refiere a cómo las plataformas de redes sociales usan sistemas automatizados – usualmente conocidos como sistemas de recomendación – para clasificar, promocionar o disminuir la visibilidad (*down ranking*) del contenido en el *feed* de noticias, normalmente basándose en los perfiles de sus usuarios. El contenido también puede ser promovido en las plataformas a cambio de un pago. Las plataformas también pueden curar contenidos a través de anuncios intersticiales – mensajes de advertencia que aparecen antes de mostrar el contenido – para advertir a los usuarios sobre cierto contenido sensible o aplicando etiquetas para destacar, por ejemplo, si el contenido proviene de una fuente confiable.

En pocas palabras, la moderación de contenidos consiste en garantizar que el contenido publicado no infrinja ninguna norma. La curación de contenidos tiene que ver con cómo el contenido se prioriza y presenta al usuario, por ejemplo, qué aparece en la parte

superior del *feed* de noticias. Ambos procesos pueden sobreponerse. Por ejemplo, bajar la categoría de cierto contenido puede ser una medida de la moderación de contenido, pero también una parte inevitable del proceso de curación de contenidos.

Términos de servicio y normas comunitarias

Compartir información u opiniones en plataformas de redes sociales no es una actividad libre de control. Cuando los usuarios se unen a Facebook, TikTok, Twitter o YouTube, aceptan cumplir con los términos de servicio de esas empresas, que rigen la relación contractual entre una empresa de redes sociales y el usuario. Esta relación contractual establece los parámetros de, por ejemplo, el acceso y uso de los diferentes productos, aplicaciones y servicios ofrecidos.

Estos términos de servicio suelen incluir las normas comunitarias de la plataforma, a veces llamadas directrices o políticas comunitarias (véanse, por ejemplo, las [normas comunitarias de Facebook](#), las [reglas y políticas de Twitter](#), o los [lineamientos de la comunidad de YouTube](#) o [TikTok](#)). Estas normas comunitarias suelen establecer los tipos de contenidos que la empresa permite o prohíbe en su plataforma. Además, pueden vetar a determinados grupos o individuos considerados peligrosos o prohibir ciertos comportamientos en línea (como la suplantación de identidad o el spam).

A los usuarios de redes sociales que incumplan estas normas se les puede retirar o disminuir la visibilidad (*down-ranking*) del contenido, o incluso inhabilitar la cuenta.

En cuanto al enfoque de este manual sobre el 'discurso del odio' y la 'desinformación', la mayoría de las normas de las distintas empresas tratan este tipo de contenido de una forma u otra, aunque con distintos grados de precisión y etiquetando el contenido de diferentes maneras.

Por ejemplo, en lo que respecta a las normas comunitarias que tratan con información 'falsa' en el sentido más amplio, el enfoque puede variar considerablemente entre plataformas.

- [Las normas de TikTok](#) sobre 'integridad y autenticidad' establecen que se eliminan contenidos o cuentas que incluyan información engañosa que cause un daño significativo. TikTok define además la 'desinformación perjudicial' como el contenido

que es inexacto o falso y que causa un daño significativo a las personas, su comunidad o el público en general, independientemente de la intención. El daño significativo puede incluir el 'menoscabo de la confianza pública en las instituciones y procesos cívicos como gobiernos, elecciones y organismos científicos', pero no incluye 'simplemente información inexacta, mitos o daños comerciales o reputacionales'.

- [Twitter centra sus acciones](#) -que van desde limitar la amplificación hasta eliminar o contextualizar- en la información 'falsa' o 'engañosa' que pueda perjudicar a las poblaciones afectadas por situaciones de crisis (como conflictos armados, emergencias de salud pública y catástrofes naturales a gran escala); contenidos engañosos (definidos como contenidos sintéticos, manipulados o descontextualizados que puedan engañar o confundir a la gente); o contenidos que pretendan manipular o interferir en elecciones u otros procesos cívicos.
- [Meta afirma](#) que elimina 'la desinformación cuando es probable que contribuya directamente al riesgo de un daño físico inminente'. Además, elimina 'contenido que es probable que contribuya de manera directa a generar una interferencia en el funcionamiento de los procesos políticos y determinado contenido multimedia manipulado y extremadamente engañoso'.

Todas estas plataformas también tienen políticas sobre el comportamiento en otras áreas que a menudo se superponen con la difusión de 'desinformación', por ejemplo, sobre cuentas falsas, fraude o comportamiento inauténtico coordinado.

Cada una de estas normas comunitarias podría ser objeto de un análisis detallado y plantear problemas y preocupaciones desde la perspectiva de la libertad de expresión (véase, por ejemplo, [el análisis que hizo ARTICLE 19 en 2018 de los términos de servicio y las normas comunitarias de Facebook, Twitter y YouTube](#)). Si bien dicho análisis supera el objetivo de este manual, en la siguiente sección presentamos una serie de preocupaciones frente al sistema actual.

Preocupaciones que plantea la regulación del discurso por contrato

La privatización de la regulación de la libertad de expresión (es decir, la regulación del discurso por contrato) suscita serias preocupaciones para la protección de la libertad de expresión. Estas preocupaciones se acentúan debido a que un número [reducido de empresas de redes sociales retienen inmenso poder sobre lo que las personas ven y comparten en línea](#), con poca transparencia y rendición de cuentas.

Reducción de los estándares de libertad de expresión

Las normas comunitarias en los términos de servicio usualmente tienen estándares más bajos para las restricciones a la libertad de expresión que los permitidos según el derecho internacional de los derechos humanos. Por ejemplo, como mencionamos en el contexto de las políticas de las empresas sobre el ‘discurso de odio’, las políticas de TikTok no permiten la misoginia. Sin embargo, una prohibición absoluta y de plano de la misoginia no satisface los requisitos de los Artículos 19 y 20 del PIDCP.

Además, en sus términos de servicio algunas empresas de redes sociales se reservan el derecho de retirar contenido a su discreción y sin motivo alguno. Por ejemplo:

- [Los Términos de Servicio de Snapchat](#) establecen: ‘Podemos restringir, finalizar o suspender temporalmente tu acceso a los servicios si no cumples con estas Condiciones, nuestras [Pautas para la comunidad](#) o la normativa legal, por motivos ajenos a nuestro control o por cualquier otro motivo’.
- [Los Términos de Servicio de TikTok](#) indican: ‘Podemos eliminar o restringir el acceso a cualquier contenido, incluido el suyo, si creemos razonablemente que (i) infringe estas Condiciones o nuestras Directrices de la comunidad; o (ii) nos perjudica a nosotros, las Filiales, nuestros usuarios o terceros. Nuestras Directrices de la comunidad establecen cómo se pueden eliminar o restringir contenidos en la Plataforma’.

ARTICLE 19 reconoce que las empresas de redes sociales son en principio libres de restringir el contenido en sus plataformas sobre la base de la libertad contractual, pero de

todas formas deberían respetar los derechos humanos, incluyendo los derechos a la libertad de expresión, privacidad y debido proceso en línea con los [Principios Rectores](#). Reservarse el derecho de remover contenido ‘por cualquier motivo o sin motivo’ claramente no está a la altura de la responsabilidad de respetar los derechos humanos.

Aunque las empresas de redes sociales legalmente tienen más margen de maniobra para restringir la libertad de expresión en sus plataformas en comparación con las restricciones que pueden imponer los estados bajo el derecho internacional de los derechos humanos, también es problemático que las normas de moderación de contenidos en estos espacios cuasi públicos no se rijan por los principios de necesidad y proporcionalidad. De hecho, a menudo parecen dictadas por el deseo de las empresas de incrementar sus ganancias y satisfacer las demandas de la industria publicitaria, que no quiere ser asociada a contenidos ofensivos, chocantes o perturbadores, aunque dicho contenido se encuentre protegido bajo los estándares de la libertad de expresión. En la práctica, los bajos estándares de protección a la libertad de expresión son por lo general el resultado de que las empresas adaptan sus normas comunitarias a los requerimientos legales nacionales, que están por debajo de los estándares internacionales de libertad de expresión²⁶.

Falta de transparencia y rendición de cuentas

Existe una considerable falta de transparencia alrededor de la implementación de las normas comunitarias. Esto repercute negativamente en la capacidad de exigir que las empresas rindan cuentas por la remoción errónea, arbitraria o discriminatoria de contenido.

En algunos casos, la falta de transparencia puede deberse al contenido mismo de una política. La política de [Personas y organizaciones peligrosas](#) de Meta ha sido un ejemplo prominente de esta cuestión. Meta designa individuos, organizaciones o redes de personas que la compañía considera que ‘proclaman una misión violenta’ o que estén ‘comprometidos con la violencia’ y remueve el contenido de apoyo a dichas entidades. Sin embargo, aunque esta política es responsable de una gran cantidad de la moderación de

contenido en las plataformas de Meta, la lista actual de entidades peligrosas no se encuentra disponible al público, [lo que hace imposible examinar su contenido](#).

Resulta además problemático que estas normas comunitarias no se encuentren disponibles por completo en varios idiomas, lo que hace imposible que muchos usuarios comprendan las reglas que rigen su discurso en línea.²⁷

En lo que respecta a los informes de transparencia, algunas empresas de redes sociales han hecho esfuerzos para mejorar sus prácticas en los últimos años. Sin embargo, sus reportes actuales sobre cómo aplican sus normas comunitarias carecen del nivel de detalle y la calidad requeridos para obtener información significativa.

Por poner solo un ejemplo ilustrativo, [Meta publica en sus reportes de transparencia información sobre la eliminación de contenido](#) basada en sus términos de servicio haciendo referencia a las áreas de su política de contenidos. Sin embargo, compartir datos agregados sobre cuántos contenidos se han retirado sobre la base de categorías tan complejas y amplias como ‘discurso de odio’ no es especialmente informativo. Falta información, entre otras, sobre cómo Meta aplica su definición de ‘discurso de odio’ (y con qué datos alimenta sus herramientas automatizadas de moderación de contenido); el número de retiradas desglosadas en distintas categorías (por ejemplo, por características protegidas) y por país (para entender si las normas comunitarias se aplican de manera diferenciada de país a país); y qué porcentaje de retiradas basadas en la política de ‘discurso de odio’ se realizan tras la notificación de las autoridades gubernamentales en distintos países. Asimismo, la mayoría de reportes de transparencia se centran en la retirada de contenido y no abordan medidas como disminución de visibilidad (*down ranking*).²⁸

Esta falta de transparencia dificulta conocer si las normas comunitarias se aplican razonable y consistentemente o de manera arbitraria y discriminatoria, a menos que haya cobertura periodística en ciertos casos o campañas públicas organizadas por grupos e individuos afectados.

El informe de 2021 de la UNESCO titulado [*Dejar entrar el sol: transparencia y responsabilidad en la era digital*](#) presenta una serie de principios de transparencia de alto nivel que podrían mejorar la transparencia de las plataformas de internet, centrándose, entre otras, en la transparencia del contenido y de los procesos, el uso y la recolección de datos personales, y la debida diligencia y reparación.

Falta de garantías procesales y recursos de reposición

Las garantías procesales que aplican a la eliminación de contenidos en las redes sociales son insuficientes. Nuevamente, la transparencia es un problema. No siempre es claro si las empresas notifican a los usuarios que su contenido fue eliminado o marcado, o si su cuenta ha sido sancionada de alguna forma, ni de las razones que sustentan estas acciones. Incluso cuando se notifica a los usuarios, la notificación suele contener una simple referencia a la política que supuestamente se ha infringido, sin una motivación suficiente que permita entender por qué se ha tomado dicha acción restrictiva de la libertad de expresión. Como se explica con más detalle a continuación, este problema se agrava cuando intervienen herramientas automatizadas para la moderación de contenidos.

Aunque la mayoría de las grandes empresas de redes sociales le permiten a los usuarios apelar la eliminación de contenido o la suspensión de cuentas, este instrumento sólo tiene sentido si los usuarios han sido notificados adecuadamente y entienden el razonamiento detrás de las sanciones que llevaron a la restricción de sus contenidos. En ese contexto, es particularmente problemático que las empresas de redes sociales hayan adoptado una práctica conocida coloquialmente como '[*shadowbanning*](#)', que describe los casos en los que se oculta o reduce la visibilidad del contenido de los usuarios sin que la plataforma les informe.

Otra falencia importante es que las personas cuyos contenidos son retirados por lo general no cuentan con recursos legales adecuados. Usualmente, los términos de servicio de las empresas no conceden ninguna base para reclamar judicialmente asuntos relacionados con la restricción del contenido y en la mayoría de las jurisdicciones los usuarios tampoco

pueden acudir a reclamos extracontractuales. Las problemáticas cláusulas sobre resolución de controversias y legislación aplicable en los términos de servicios de las empresas de redes sociales pueden crear barreras adicionales en el acceso a la justicia de los usuarios. En algunos casos, los términos de servicio pueden impedir que los usuarios presenten reclamaciones en las cortes locales de sus países de residencia o que apliquen la legislación local a los términos de servicio. Esto disuade a la mayoría de los usuarios de iniciar un litigio, ya que carecen de los recursos para hacerlo.

Por ejemplo, para usuarios residentes en el Reino Unido – cláusulas similares pueden aplicar en otras jurisdicciones – [Twitter establece](#) que todas las disputas relacionadas con sus términos de servicio deben presentarse ante los tribunales de San Francisco y que la legislación aplicable es la del Estado de California. Los [términos de servicio de Snapchat](#) contienen una cláusula compromisoria según la cual se debe realizar un arbitraje en Estados Unidos; además, contienen una renuncia a las acciones colectivas (excepto si es aplicable para procedimientos de tribunales de pequeños reclamos y con una opción de darse de baja (*opt-out*) de 30 días, que la mayoría de usuarios nunca notarán). Los [términos de Meta](#) son más razonables en tanto establecen los tribunales del estado donde resida el usuario serán competentes para conocer de disputas sobre asuntos de consumo y se aplicará la legislación de dicho estado. Los [términos de servicio de TikTok](#) estipulan que las disputas relacionadas con dichos términos se someterán a la jurisdicción de los tribunales nacionales del usuario, así como a los de las cortes de la República de Irlanda, de Inglaterra y de Gales.

Elusión del Estado de Derecho

Por último, las autoridades públicas, y en especial los cuerpos y agencias de seguridad, solicitan periódicamente la cooperación de las plataformas de redes sociales con el fin de combatir la actividad criminal (por ejemplo, la difusión de contenido relativo al abuso sexual a menores) u otros daños sociales (por ejemplo, el ‘extremismo en línea’) de [formas que eluden el Estado de Derecho](#). En particular, dado que las autoridades no siempre están facultadas para ordenar la eliminación del contenido en cuestión, en ocasiones contactan

a las empresas de redes sociales de manera informal y solicitan la eliminación del contenido basándose en los términos de servicio de las empresas. Si bien las empresas por lo general no se encuentran legalmente obligadas a cumplir con dichas solicitudes, se encuentran en una situación difícil, particularmente en escenarios en los que el contenido puede estar en el borde de la ilegalidad. El resultado final es que las empresas de redes sociales se convierten, con frecuencia, en el largo brazo de la ley sin que los usuarios tengan la oportunidad de impugnar la legalidad de las restricciones a la libertad de expresión ante los tribunales²⁹.

Ejemplo

En el caso de la 'música drill británica', el Consejo Asesor de Contenidos anuló la decisión de Meta de remover un video musical – 'Secrets Not Safe' de Chinx (OS) – de Instagram. Meta había eliminado el contenido inicialmente tras una solicitud de la Policía Metropolitana del Reino Unido. Esta autoridad había enviado un correo electrónico a Meta solicitando que la empresa revisara el contenido que incluyera 'Secrets Not Safe' y había proporcionado contexto adicional sobre la violencia de pandillas en Londres, incluyendo asesinatos, y la preocupación de que la pista diera lugar a más represalias. Meta removi  el contenido de la cuenta en cuesti n por violar su pol tica sobre violencia e incitaci n.

El Consejo Asesor de Contenidos encontr , entre otras cosas, que '[l]os canales a trav s de los cuales las autoridades policiales presentan solicitudes a Meta son aleatorios y confusos. No se exige a las autoridades policiales cumplir con criterios m nimos para justificar sus solicitudes y, por lo tanto, las interacciones no son sistem ticas. Asimismo, los datos que Meta publica sobre las solicitudes gubernamentales son parciales.'

El papel de los marcos regulatorios

Los procesos de moderaci n de contenido no solo deben hacer cumplir los t rminos de servicio y normas comunitarias de las empresas de redes sociales, sino que tambi n deben hacer cumplir las obligaciones normativas de remover contenidos ilegales o reprochables. En los  ltimos a os, los gobiernos han incrementado la presi n a las empresas para retirar m s contenido de sus plataformas – desde 'discurso de odio' y 'extremista' hasta 'desinformaci n'.

Enfoque tradicional en la regulación de la responsabilidad de intermediarios

Tradicionalmente, los marcos regulatorios se han centrado en regular la responsabilidad de los intermediarios de internet – un término amplio que incluye a las empresas de alojamiento web, a los proveedores del servicio de internet, a los motores de búsqueda y a las plataformas de redes sociales³⁰. La legislación que regula la responsabilidad de los intermediarios determina hasta qué punto los intermediarios de internet son legalmente responsables – y si están obligados, por ejemplo, a indemnizar a una parte agraviada– por el contenido difundido o creado por sus usuarios (contenido de terceros).

En términos generales, [los regímenes de responsabilidad oscilan desde la responsabilidad estricta, en un extremo del espectro, y la inmunidad, en el otro](#). Bajo los regímenes de responsabilidad estricta, los intermediarios de internet pueden ser demandados ante los tribunales por la conducta indebida de sus usuarios sin la necesidad de que exista culpa o conocimiento por parte del intermediario. Para cumplir con la ley, en la práctica los intermediarios monitorean el contenido y toman medidas cuando lo consideran relevante. Este modelo ha sido aplicado en Tailandia.

Ejemplo

Chiranuch Premchaiporn, el editor de Prachatai, un portal de noticias en línea tailandés, fue juzgado y sentenciado bajo las reglas de la Ley tailandesa de Delitos Informáticos de 2007 por no haber retirado de manera expedita un comentario anónimo que fue considerado como ofensivo hacia el Rey³¹. Esta Ley castiga los ‘datos falsos’ que ocasionan daño a terceras personas, causan pánico público o socavan la defensa nacional, así como ‘cualquier proveedor del servicio que intencionalmente soporte’ dichos datos falsos. El código penal de Tailandia (*lèse-majesté*) establece que cualquiera que ‘difame, insulte o amenace al rey, la reina, el heredero o el regente’ será sentenciado a prisión.

Chiranuch Premchaiporn fue sentenciado a un año de prisión con una multa de 30,000 baht, que luego se redujo a ocho meses de suspensión y una multa de 20,000 baht luego de la apelación.³²

La inmunidad absoluta frente al contenido generado por los usuarios – es decir, que no es posible interponer una demanda contra los intermediarios basada en el contenido generado por los usuarios – es poco común. El caso más destacado que promueve este enfoque es la Sección 230 de la [Communications Decency Act](#) [Ley de Decencia de las Comunicaciones] de 1996, una legislación que es aplicable en los Estados Unidos. La Sección 230 otorga inmunidad legal a las plataformas de internet por los contenidos publicados por terceros (aunque no se extiende a la inmunidad por violaciones a la legislación federal penal, la legislación sobre propiedad intelectual, o la legislación sobre privacidad de las comunicaciones electrónicas).

Muchos sistemas jurídicos se encuentran entre la responsabilidad objetiva y la inmunidad absoluta adoptando un sistema de inmunidad condicional. Con frecuencia, los intermediarios de Internet se eximen de responsabilidad siempre que retiren los contenidos una vez tengan *conocimiento real* de la ilegalidad. Estos sistemas de responsabilidad basados en el conocimiento generalmente operan mediante procedimiento de 'notificación y retirada'.

El funcionamiento exacto de estos procedimientos de 'notificación y retirada' varía entre jurisdicciones. Por lo general, se considera que un intermediario de Internet adquiere conocimiento de la naturaleza ilegal de un contenido una vez que es notificado por un tercero. Si el intermediario de Internet no elimina el contenido ilegal a pesar de dicha notificación, puede ser considerado legalmente responsable por cualquier daño causado. Por ejemplo, la recién adoptada [Ley de Servicios Digitales de la Unión Europea](#) establece la responsabilidad de los intermediarios sobre los contenidos que hayan sido objeto de notificaciones debidamente motivadas por los usuarios

Desde una perspectiva de libertad de expresión, es ampliamente reconocido –incluso por los mandatos especiales sobre libertad de expresión– que la inmunidad ampliada de responsabilidad de los intermediarios de Internet es una de las formas más eficaces de proteger la libertad de expresión en línea. Si las empresas pueden asumir la responsabilidad de los contenidos publicados por sus usuarios, en la práctica se ven obligadas a monitorear

todos los contenidos generados por los usuarios, lo que supone una enorme invasión de su derecho a la privacidad. Un régimen de este tipo también genera fuertes incentivos para que las empresas censuren en exceso a sus usuarios y retiren contenido que puede ser perfectamente legal para evitar cualquier riesgo de infringir la ley.

La experiencia demuestra que incluso los regímenes de inmunidad condicional que operan a través de procedimientos de 'notificación y retirada' incentivan la pronta remoción de contenido con base en las acusaciones de un particular o un organismo público, sin que se determine judicialmente si el contenido en cuestión es ilegal. Además, la persona que publicó el contenido en cuestión no suele tener la oportunidad de examinar la denuncia.

En 2011, Frank La Rue, ex Relator Especial de la ONU para la Libertad de Expresión, afirmó que la censura nunca debe delegarse a una entidad privada y que los Estados no deben utilizar ni obligar a los intermediarios a ejercer la censura en su nombre.³³ También señaló que los regímenes de 'notificación y retirada' estaban expuestos a abusos tanto por parte de los Estados como de actores privados y que la falta de transparencia en relación con la toma de decisiones por parte de los intermediarios a menudo ocultaba prácticas discriminatorias o presiones políticas que afectaban las decisiones de las empresas.³⁴

Tendencia hacia una mayor regulación de las plataformas en línea

En los últimos años, las empresas de redes sociales han sido cada vez más criticadas por incrementar sus ganancias por medio de algoritmos que promueven un *engagement* adictivo con contenidos 'extremistas' y otros contenidos 'perjudiciales'. Esto ha planteado la pregunta de si se necesita una mayor regulación para controlar el poder de las principales empresas de redes sociales, atajar el contenido ilegal y perjudicial, y garantizar mayor responsabilidad democrática al público en general por sus decisiones. Los gobiernos han respondido con propuestas según las cuales las plataformas tienen un 'deber de protección' hacia sus usuarios para prevenir los 'daños' causado por el discurso de otros usuarios de la plataforma.³⁵

Muchas de las propuestas actuales de marcos regulatorios son profundamente problemáticas desde el punto de vista de la libertad de expresión. Aunque aparentemente buscan aumentar la rendición de cuentas por parte de las empresas de redes sociales, usualmente se centran en la regulación del 'contenido' en línea. Esto significa que, en la práctica, regulan el discurso de los usuarios en lugar de los productos, sistemas y procesos aplicados por las empresas de redes sociales. Los Estados exigen efectivamente que las empresas monitoreen las comunicaciones humanas y decidan qué discurso es 'ilegal' o 'perjudicial',³⁶ cuando en realidad debería ser responsabilidad de autoridades judiciales independientes tomar esas decisiones.

Además de las preocupaciones legítimas sobre la externalización de decisiones sobre la legalidad del discurso de usuarios a actores privados, en la mayoría de los casos estas evaluaciones son extremadamente complejas y dependientes del contexto y, por lo tanto, deberían ser realizadas por individuos capacitados. Sin embargo, la realidad es que las empresas de redes sociales utilizan sistemas de moderación algorítmica, como herramientas automatizadas de coincidencia de *hash* y de *machine-learning*, para moderar los contenidos. Dado que estas tecnologías actualmente no están lo suficientemente avanzadas (y es posible que nunca lo estén) para distinguir de manera confiable entre contenidos legales e ilegales, habitualmente identifican contenido legal como ilegal y eliminan cantidades considerables de contenido legítimo.

La regulación debe adoptar un enfoque basado en los derechos humanos e incluir los mercados digitales

En lugar de obligar a las plataformas a restringir discursos indeseables de los usuarios, las propuestas deberían garantizar que los derechos humanos estén en el centro de la regulación de las plataformas. Los principios de legalidad, legitimidad, necesidad y proporcionalidad establecidos en el artículo 19(3) del ICCPR deben aplicarse en todo momento. Como cualquier marco que imponga limitaciones a la libertad de expresión, la regulación que rige a las empresas de redes sociales debe basarse en evidencia sólida y priorizar las medidas menos lesivas y restrictivas para abordar los daños en línea.

¿Qué significa esto en la práctica? En lugar de pedir a las plataformas ejercer aún más controles sobre el discurso de los usuarios mediante la revisión y evaluación de todo el contenido que generan, los reguladores deberían centrarse en métodos menos intrusivos que estén específicamente diseñados para abordar algunos de los efectos negativos de los modelos comerciales usados por las empresas de redes sociales, incluidos sus sistemas de recomendación. Por ejemplo, las soluciones regulatorias deberían exigir mayor transparencia de las empresas con los reguladores, investigadores y usuarios sobre el funcionamiento de sus sistemas de recomendación; establecer límites claros sobre la cantidad de datos de los usuarios que se les permite recopilar; y exigir la implementación de procesos de debida diligencia en materia de derechos humanos. También deberían centrarse en la transparencia sobre las decisiones de moderación de contenidos y en mejorar los sistemas para resolver cualquier disputa que estas decisiones provoquen.

[ARTICLE 19 también ha abogado](#) durante años porque las soluciones regulatorias aborden además la posición dominante de las principales plataformas en línea mediante herramientas regulatorias que aumenten la competencia en el mercado y mejoren la capacidad de elección de los usuarios sobre los contenidos que pueden ver en línea.

Algunas de estas soluciones regulatorias fueron adoptadas en la UE en 2022 mediante la Ley de Servicios Digitales de la UE y la [Ley de Mercados Digitales](#). Si bien estos marcos regulatorios podrían haber sido más ambiciosos en cuanto a la protección de los derechos humanos en línea (por ejemplo, estableciendo un derecho explícito para los usuarios al cifrado y al anonimato), se centran acertadamente en reequilibrar los mercados digitales y regular los sistemas de moderación y curación de contenidos utilizados por las empresas de redes sociales.

Asimismo, varias de las normas de la Ley de Servicios Digitales de la UE sobre moderación de contenidos abordan algunas de las inquietudes suscitadas por la 'regulación de la expresión por contrato'. Entre otras cosas, exigen:

- que se notifique a los usuarios y se les proporcione una exposición de motivos en caso de que se restrinja la visibilidad de sus contenidos (incluida la eliminación o relegación);
- el establecimiento de un mecanismo interno de reclamación, que permita a los usuarios presentar quejas contra las decisiones de moderación de contenidos a la empresa electrónicamente y de forma gratuita;
- que los Estados miembros establezcan un mecanismo de resolución extrajudicial de litigios sobre decisiones de moderación de contenidos, lo que significa que los usuarios tienen derecho a elegir cualquier órgano de resolución extrajudicial de litigios para resolver disputas relacionadas con decisiones de moderación de contenidos, incluidas las que no pudieron resolverse mediante el sistema interno de gestión de reclamaciones; y
- informes de transparencia relativamente detallados sobre la moderación de contenidos que lleven a cabo las empresas de redes sociales; por ejemplo, sobre el número de órdenes recibidas de las autoridades de los Estados miembros para actuar contra los contenidos ilícitos o la moderación de contenidos llevada a cabo por iniciativa propia de las empresas de redes sociales, incluido el número y tipo de medidas adoptadas que afecten la disponibilidad, visibilidad y accesibilidad de los contenidos publicados.

La Ley de Servicios Digitales podría, hasta cierto punto, servir como modelo para propuestas regulatorias en todo el mundo. Un ejemplo claro es la propuesta de [Ley de Libertad, Responsabilidad y Transparencia](#) que se está discutiendo en Brasil durante la redacción de este documento. Anteriormente, la Ley alemana de Redes (NetzDG), que obliga a las plataformas a retirar contenidos 'manifiestamente ilegales' dentro de las 24 horas posteriores a la notificación bajo la amenaza de fuertes multas, ya había inspirado a varios Estados, entre ellos Kenia, Malasia e India, a reforzar las leyes de responsabilidad de intermediarios.³⁷

En los foros internacionales también se está discutiendo cómo abordar la regulación de las plataformas de forma respetuosa con los derechos humanos y cómo hacer frente a las oportunidades y riesgos para los derechos humanos derivados de las tecnologías digitales en general. Por ejemplo, en las Naciones Unidas se están realizando esfuerzos para acordar

un [Pacto Digital Mundial](#) que aborde una serie de cuestiones, entre ellas 'promover la confianza en Internet introduciendo criterios de rendición de cuentas por la discriminación y los contenidos engañosos'. [La UNESCO está elaborando directrices globales](#) para regular las plataformas digitales, con el objetivo de 'forma[r] parte de los procesos regulatorios que estén en desarrollo o revisión respecto a las plataformas digitales, de modo coherente con la normativa internacional en materia de derechos humanos'.

Medios de comunicación y moderación de contenidos

La protección del pluralismo y la diversidad de medios en el ecosistema digital ha sido motivo de inquietud para los reguladores. Los medios de comunicación dependen de las plataformas para acceder a sus audiencias, ingresos publicitarios y financiamiento, lo que lleva a una influencia significativa de las plataformas sobre sus decisiones editoriales, organizativas y comerciales. Específicamente, ha suscitado inquietud la influencia de los sistemas de moderación de contenidos de las plataformas en la visibilidad, la monetización y el alcance de los contenidos y las cuentas, así como [su impacto en la independencia editorial de los periodistas](#) y la sostenibilidad financiera del periodismo.

Las deficiencias de los sistemas automatizados de moderación de contenidos –que se analizan con más detalle en la siguiente sección de este manual – han complicado aún más las cosas para los medios de comunicación. Debido a su incapacidad para considerar adecuadamente el contexto, las herramientas automatizadas también tienden a eliminar información de interés público, como, por ejemplo, información sobre grupos extremistas o violaciones de los derechos humanos.

Ejemplo

A un medio de comunicación de Sarajevo se le impidió publicar en Facebook contenidos que informaban sobre la sentencia del Tribunal Penal Internacional para Yugoslavia (TPIY) en el caso de Ratko Mladić, condenado por varios crímenes internacionales, algunos relacionados con la masacre de Srebrenica. Esto se debió a que Facebook había etiquetado erróneamente al TPIY como una organización criminal. Por lo tanto, los sistemas automatizados de moderación de contenidos concluyeron que el artículo sobre el juicio de Mladić pretendía promover una 'organización criminal'.³⁸

Algunos marcos regulatorios propuestos incluyen disposiciones que protegerían los 'contenidos editoriales' de las reglas de moderación de contenidos (mientras que otras propuestas regulatorias, como las de [Australia](#) o [Canadá](#), se han centrado más en obligar a las empresas de redes sociales a pagar por las noticias que utilizan). Por ejemplo, el [Proyecto de Ley de Seguridad en Línea](#) del Reino Unido –en negociación durante la redacción de este documento– excluye los contenidos editoriales de noticias de sus reglas de moderación de contenido, excluyéndolos del ámbito de las obligaciones que el proyecto de ley impone a las empresas de redes sociales. Una exención similar se debatió durante las negociaciones de la Ley de Servicios Digitales de la UE, pero finalmente fue rechazada. La [Ley Europea de Libertad de los Medios de Comunicación](#), también en negociación durante la redacción de este documento, propone un proceso mediante el cual los prestadores de servicios de medios de comunicación podrían solicitar a las plataformas un trato especial con respecto a la manera en que se moderan sus contenidos.³⁹ Más concretamente, la propuesta actual establece que, antes de suspender cualquier contenido, las empresas de redes sociales tendrían que comunicar el motivo de su decisión al medio de comunicación y garantizar que las denuncias del proveedor 'se tramiten y se resuelvan de manera prioritaria y sin demora indebida'.⁴⁰

La pertinencia de crear normas especiales de moderación de contenidos para los prestadores de servicios de medios de comunicación es objeto de controversia. Hay quienes sostienen que los términos de servicio de las empresas de redes sociales no deben prevalecer sobre los propios criterios editoriales de los medios de comunicación y que los actores de los medios suelen responder ante la ley, las normas éticas y la pertenencia a asociaciones profesionales.

Sin embargo, organizaciones dedicadas a la defensa de la libertad de expresión y los derechos digitales, entre ellas ARTICLE 19, han sostenido que, a pesar de los indiscutibles retos a los que se enfrentan los actores de los medios para navegar por las normas de moderación de contenidos de las plataformas – y el hecho de que la asimetría de poder entre las empresas de redes sociales y los actores de los medios requiere soluciones

regulatorias – es problemático desde la perspectiva de la libertad de expresión conceder un trato especial a ciertos actores en el mercado de los medios de comunicación cuando se trata de la moderación de sus contenidos. Por principio, el discurso de algunos actores no debería valorarse más que el de otros en virtud de quiénes son en lugar de por lo que dicen.

Teniendo en cuenta los estándares de libertad de expresión sobre 'discurso de odio', en particular, como se explicó anteriormente, la posición e influencia del orador es uno de los factores clave que los tribunales deben considerar al evaluar si el discurso alcanza el nivel de incitación al odio prohibido por el Artículo 20(2) del PIDCP. Por lo tanto, pedir a las empresas de redes sociales que eliminen el contenido de odio publicado por usuarios particulares, pero que protejan el mismo contenido si es publicado por actores en el mercado de los medios de comunicación exentos iría en contra de estos estándares.

También es probable que las excepciones para los medios de comunicación refuercen el poder de los titulares en detrimento de periodistas ciudadanos, blogueros más pequeños o activistas que no cumplen con los criterios para acogerse a las exenciones regulatorias, incluso cuando ejercen una actividad periodística sin ánimo de lucro. Por ello, las organizaciones de la libertad de expresión rechazan en general cualquier tipo de privilegio para los medios de comunicación en materia de moderación de contenidos.

Procesos de moderación de contenidos

Las empresas de redes sociales emplean diversos enfoques en la moderación de contenidos y usan diversos tipos de herramientas para hacer cumplir sus políticas y restringir o eliminar contenidos y cuentas ilegales o reprochables. Dado el volumen de contenido producido por los usuarios – las cifras más recientes muestran, por ejemplo, que cada minuto al menos [350.000 trinos se comparten en Twitter](#), [500 horas de video se suben a YouTube](#) y [más de 510.000 comentarios y 136.000 fotos se comparten en Facebook](#) – la mayoría de las empresas de redes sociales más grandes han optado por acudir en gran

medida a herramientas automatizadas para reducir la moderación humana que toma más tiempo, reservando esta para casos específicos de revisión de contenidos.

Las deficiencias de la automatización

La moderación de contenidos automatizada implica el uso de herramientas automatizadas de detección, filtrado y moderación para marcar, separar y eliminar contenidos y cuentas específicas. Existen multitud de herramientas automatizadas, muchas de ellas basadas en inteligencia artificial y *machine-learning*, que se pueden usar durante el proceso de moderación de contenidos. Estas herramientas pueden desplegarse en una amplia gama de categorías de contenido y formatos multimedia en diferentes etapas del ciclo de vida del contenido para identificar, clasificar y eliminar contenidos. Algunos de los métodos y herramientas de automatización más usados son la tecnología de *hash* digital,⁴¹ el reconocimiento de imágenes,⁴² o el procesamiento de lenguaje natural *natural language processing* (NLP por sus siglas en inglés).⁴³

Las plataformas suelen aplicar la llamada moderación de contenidos híbrida, la cual implica el uso de herramientas automatizadas para marcar y priorizar ciertos casos para que luego sean revisores humanos los que emitan una decisión final sobre el contenido en cuestión.⁴⁴

Solo es posible que las plataformas moderen contenidos a gran escala si confían hasta cierto punto en la moderación de contenidos automatizada, en tanto la moderación humana es inviable al ritmo en que los usuarios generan nueva información. Al mismo tiempo, estas herramientas entrañan serios riesgos desde la perspectiva de la libertad de expresión, en particular cuando se aplican categorías complejas como contenido relativo a ‘discurso de odio’ o ‘terrorismo’. Dichas categorías requieren un nivel de análisis contextual que comprenda los matices que estas herramientas no pueden proporcionar.

Falta de precisión y fiabilidad

La precisión de una herramienta automatizada de moderación de contenidos para detectar y eliminar adecuadamente el contenido depende, en gran medida, del contenido para el cual

esté entrenada a enfrentar. Por ejemplo, las herramientas automatizadas pueden ser efectivas a la hora de identificar contenidos calificados como material de abuso sexual infantil. En este caso, existe amplio consenso internacional de que se trata de contenido ilegal, hay parámetros claros de lo que debe marcarse y los modelos se han entrenado con suficientes datos para ofrecer altos niveles de precisión⁴⁵.

Este no es el caso cuando se trata de categorías como contenido de ‘discursos de odio’ o ‘terrorista’. Para que los clasificadores del NLP funcionen con precisión, necesitan parámetros y definiciones del discurso⁴⁶. Pero las definiciones de contenido extremista o terrorista – ni hablar de qué constituye exactamente elogio o apoyo a una organización terrorista – son inexistentes o notoriamente vagas. De manera más general, las herramientas de NLP por lo general no son capaces de comprender los matices y elementos contextuales del discurso o de identificar cuándo un contenido es una sátira o se publica con fines informativos⁴⁷. Es fundamental que periodistas y organizaciones de derechos humanos puedan generar conciencia sobre las atrocidades terroristas: filtrar y eliminar ‘contenido terrorista’ sin apreciar el contexto pone en peligro el cubrimiento periodístico legítimo y la documentación de violaciones de derechos humanos. Existen múltiples casos en los que la moderación automatizada ha dado lugar a la eliminación excesiva de contenidos de interés público.

Ejemplo

El Archivo Sirio ([*Syrian Archive*](#)), un proyecto que tiene como objetivo preservar evidencia de violaciones de derechos humanos y otros crímenes cometidos durante el conflicto en Siria con fines de defensa, justicia y rendición de cuentas, ha descubierto que se han eliminado de YouTube videos que documentan crímenes de guerra. Esto puede conducir a la pérdida generalizada y a veces permanente de lo que podría ser evidencia crucial de crímenes de guerra.

Ejemplo

En mayo de 2021 la Unión Americana de Libertades Civiles (ACLU por sus siglas en inglés) reportó que decenas de activistas y periodistas tunecinos, sirios y palestinos que cubrían abusos contra los derechos humanos y ataques aéreos contra civiles se quejaron de que Facebook había desactivado sus cuentas en virtud de su política contra 'elogio, apoyo o representación de grupos terroristas', señalando los problemas ocasionados por los 'imprecisos sistemas automatizados de detección' y por la inexistencia de una definición internacional para términos como 'terrorismo', 'extremismo violento' o 'extremismo', por no hablar de 'apoyo' a los mismos.

La precisión de los clasificadores de NLP disminuye especialmente cuando se aplican en diferentes idiomas y contextos. De hecho, las herramientas automatizadas tienen una capacidad limitada para analizar y comprender las diferencias en lenguaje y comportamiento que son consecuencia de distintos factores demográficos y regionales⁴⁸. Además, la mayoría de herramientas de NLP son menos precisas cuando analizan textos que no se encuentran en inglés debido a la falta de recursos en otros idiomas⁴⁹.

Amplificación de sesgos

Más allá del hecho de que disminuyen su precisión cuando analizan lenguajes distintos al inglés, la presencia de sesgos en herramientas automatizadas genera el riesgo de incrementar la marginalización y censura en la que se encuentran ciertos grupos, que ya deben soportar los prejuicios y la discriminación dentro y fuera de línea. Este riesgo tiene su origen en los muchos tipos de sesgos humanos que se introducen en los datos de entrenamiento y que, por tanto, pueden amplificarse mediante el uso de herramientas automatizadas⁵⁰. Si los conjuntos de datos de entrenamiento no son lo suficientemente representativos, se corre el riesgo de que los sistemas de inteligencia artificial (IA) aprendan y perpetúen cualquier sesgo subyacente en los datos⁵¹.

Ejemplo

La Unión Americana de Libertades Civiles (ACLU por sus siglas en inglés) reportó que, durante años, Meta trató el discurso de las mujeres y personas de color de forma distinta al de los hombres y personas blancas – incluso cuando se describían sus experiencias de acoso sexual y racismo. En 2017, cuando algunas mujeres de color y algunas personas blancas publicaron exactamente el mismo contenido, Meta suspendió únicamente las cuentas de mujeres de color.

Falta de transparencia y rendición de cuentas

Se necesita más transparencia y rendición de cuentas en torno al uso de herramientas automatizadas. No hay transparencia sobre cómo los conjuntos de datos se compilan, qué tan precisas son las herramientas automatizadas de moderación de contenidos y qué tanto contenido es eliminado, tanto correcta como incorrectamente⁵². Esto despierta preocupación por el derecho a la libertad de expresión de las personas cuyos contenidos han sido marcados por error o cuyas cuentas han sido erróneamente desactivadas. Estas preocupaciones se agravan aún más si se utilizan filtros al cargar contenidos y el contenido marcado desaparece incluso antes de que se publique, haciendo más difícil saber si el contenido fue eliminado por error⁵³.

Denuncias de usuarios y ‘marcadores de confianza’

Por lo general, las empresas de redes sociales permiten a sus usuarios reportar contenido que consideren que es ilegal, vulnera las normas comunitarias u ocasione daño. Como se mencionó anteriormente, algunos sistemas regulatorios atribuyen consecuencias jurídicas, en términos de responsabilidad de intermediarios, a aquellos reportes a través de procedimientos de ‘notificación y retirada’, lo cual es problemático desde una perspectiva de libertad de expresión y aumenta el riesgo de notificaciones vejatorias o abusivas.

Asimismo, algunas empresas de redes sociales acuden a un sistema de marcadores de confianza o ‘*trusted-flagger*’, en el que los reportes realizados por *trusted flaggers* son revisados más rápidamente. Normalmente, los *trusted flaggers* son personas o entidades

expertas en identificar y marcar contenido ilegal. Puede tratarse de actores de la sociedad civil.

Hoy en día, las empresas de redes sociales por lo general no proporcionan información suficiente sobre el sistema de '*trusted-flagger*'. Esto incluye cómo se selecciona a los *trusted flaggers* y en qué medida el contenido marcado por ellos se sujeta a una revisión adecuada o si, en cambio, se elimina automáticamente. Aunque el sistema de '*trusted-flagger*' puede incidir positivamente en mejorar la calidad de las notificaciones, esto no es equivalente a una evaluación imparcial o independiente del contenido en cuestión. Los *trusted flaggers* por lo general se seleccionan en virtud de su experticia en el impacto de ciertos tipos de contenidos, como los derechos de autor, contenidos relacionados con el terrorismo, 'discurso de odio', y su proximidad a las víctimas de este tipo de discurso, pero no sobre la base de su experticia en libertad de expresión. Por ende, no se encuentran necesariamente capacitados para realizar análisis imparciales sobre si la restricción del contenido en cuestión es acorde al derecho internacional de los derechos humanos⁵⁴.

La Ley de Servicios Digitales de la Unión Europea ha incorporado este sistema, dándole especial privilegio a los *trusted flaggers* y estableciendo que una vez el contenido ha sido notificado por ellos, las plataformas deben remover el contenido ilegal 'inmediatamente'. La [sociedad civil ha criticado](#) esta legislación en tanto permite que agencias gubernamentales y fuerzas de seguridad obtengan el estatus de *trusted flaggers*, lo cual podría abrir la puerta al abuso.

La necesidad de una comprensión más profunda del contexto

Las deficiencias de los sistemas automatizados de moderación de contenidos hacen indispensable que las empresas de redes sociales contraten suficientes revisores humanos e inviertan recursos para comprender mejor el contexto en el que se dan distintas formas de expresión. Esto es necesario para revisar y analizar de manera adecuada las categorías del discurso que no son susceptibles de ser analizadas por la moderación automatizada y para analizar de mejor manera las decisiones de moderación de contenidos que los

usuarios recurren. En específico, los revisores humanos deben ser entrenados y seleccionados adecuadamente para reducir el riesgo de sesgos en su toma de decisiones. También deben ser hablantes nativos de los lenguajes que les son asignados y tener conocimiento suficiente del contexto local. Por ahora, existe una significativa falta de transparencia sobre el número de revisores humanos de contenido que emplea cada empresa de redes sociales, sobre cómo las empresas asignan tareas de moderación de contenidos por país e idioma, sobre cómo son entrenados los revisores, las cuestiones específicas que deben resolver y dónde se encuentran. Esta opacidad apunta a un asunto más generalizado que tiene que ver con los pocos recursos que las grandes empresas de redes sociales invierten en entender el contenido en muchos países donde operan, pero que no consideran de importancia estratégica.

Conclusión

La moderación de contenidos es un campo en rápida evolución. Los entornos normativos cambian constantemente, nacen nuevos actores, las empresas de redes sociales actualizan sus políticas regularmente y las organizaciones internacionales diseñan iniciativas para encontrar soluciones globales o regionales al contenido ‘problemático’ en línea. La moderación de contenidos también necesita responder a los retos del mundo real, como la pandemia de Covid-19, los conflictos armados y el desarrollo de nuevos productos y tecnologías como el metaverso.

A pesar de estos cambios rápidos, algunos aspectos de la moderación de contenidos no cambiarán en un futuro cercano. Los derechos humanos deben estar en el centro de cualquier intento de regular a las empresas de redes sociales por parte de organizaciones internacionales o los Estados. Las empresas de redes sociales deben tomarse en serio su responsabilidad de respetar los derechos humanos, incluyendo cuando realizan su labor de moderación de contenidos, y de entender el contexto en el que operan.

Hasta ahora no lo han hecho. La falta de inversión para comprender de mejor manera los contextos locales incluye el fracaso a la hora de aliarse con la sociedad civil local para

obtener dicha comprensión. Las coaliciones locales sobre libertad de expresión y moderación de contenidos pueden acortar la distancia entre los actores locales y las empresas de redes sociales y garantizar que las particularidades de los contextos y un enfoque de derechos humanos sean incorporados a las decisiones de moderación de contenidos. Pero esto requiere compromisos serios de las empresas de redes sociales de respetar los derechos humanos y mitigar los posibles impactos negativos que resulten de las deficiencias en sus sistemas y procesos de moderación de contenidos.

El informe de ARTICLE 19 titulado [*Content Moderation and Freedom of Expression*](#) [Moderación de contenidos y libertad de expresión] desarrolla estos asuntos en más detalle y resume lo que está en juego: 'Cuando estas empresas no tienen en cuenta las diversas dimensiones (lingüísticas, políticas, sociales, culturales y económicas) de los contextos locales, los procesos de moderación de contenidos pueden tener impactos dramáticos en las sociedades afectadas, como el aumento de la polarización y el riesgo de violencia'.

Bibliografía

ARTICLE 19, [*'Discurso de odio': Manual*](#), 2015.

ARTICLE 19, [*Online Harassment and Abuse against Women Journalists and Major Social Media Platforms*](#) [Acoso en línea y abuso contra mujeres periodistas y grandes plataformas de redes sociales], 2020.

ARTICLE 19, [*Side-Stepping Rights: Regulating Speech by Contract. Policy Brief*](#) [Eludiendo los derechos: La regulación del discurso por contrato], 2018.

ARTICLE 19, [*Vigilar a los vigilante: Moderación de contenido, gobernanza y libertad de expresión*](#), 2023.

UNESCO, [*Hacer frente al discurso de odio en las redes sociales: desafíos contemporáneos*](#), 2021.

UNESCO, [*'Contrarrestar el discurso de odio'*](#).

UNESCO, [*Encontrar fondos para que el periodismo prospere : opciones de política para respaldar la viabilidad de los medios de comunicación*](#), 2022.

UNESCO, [*'¿Cómo abordar el discurso de odio en línea con un enfoque basado en los derechos humanos?'*](#), video de YouTube, 2022.

UNESCO, [*'Los límites legítimos a la libertad de expresión: El Plan de Acción de Rabat'*](#), video de YouTube, 2022.

UNESCO, [*Salvaguardar la libertad de expresión y el acceso a la información: directrices para un enfoque de múltiples partes interesadas en el contexto de la regulación de las plataformas digitales*](#), 2023.

UNESCO, [*'Directrices para la gobernanza de las plataformas digitales'*](#), 2023.

UNESCO, [*Declaración de Windhoek + 30: la información como bien común*](#), Día Mundial de la Libertad de Prensa, 2021.

Notas finales

¹ Los términos ‘discurso de odio’ y ‘desinformación’ no están definidos en el derecho internacional. Por esta razón, ARTICLE 19 usa los términos en comillas invertidas en toda la publicación.

² La DUDH no es vinculante para los Estados en tanto fue adoptada por medio de una resolución de la Asamblea General de las Naciones Unidas. Sin embargo, desde su adopción en 1948 varias de sus disposiciones han adquirido fuerza vinculante como costumbre del derecho internacional. Véase el caso *Filartiga v. Pena-Irala*, 630 F. 2d 876 (1980) de la Corte de Apelaciones del Segundo Circuito de los Estados Unidos.

³ Asamblea General de las Naciones Unidas, [Pacto Internacional de Derechos Civiles y Políticos](#), 16 de diciembre de 1966, Serie de Tratados, vol. 999, p. 171.

⁴ Artículo 10 del [Convenio Europeo de Derechos Humanos](#), 4 de septiembre de 1950; artículo 9 de la [Carta Africana sobre los Derechos Humanos y de los Pueblos](#) (Carta de Banjul), 27 de julio de 1981; artículo 13 de la [Convención Americana sobre Derechos Humanos](#), 22 de noviembre de 1969.

⁵ Véase Tribunal Europeo de Derechos Humanos, *Dink v. Turkey*, párrafos 106 y 137 (Aplicaciones no. 2668/07, 6102/08, 30079/08, 7072/09, y 7124/09), 14 de septiembre de 2010.

⁶ A. Kuczerawy (2017) ‘[The Power of Positive Thinking, Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression](#)’ [El poder del pensamiento positivo, la responsabilidad de los intermediarios y el disfrute efectivo del derecho a la libertad de expresión], JIPITEC, p. 226.

⁷ Comité de Derechos Humanos de la ONU, [Observación General N° 34. Artículo 19: Libertad de opinión y libertad de expresión](#), CCPR/C/GC/34, 12 de septiembre de 2011, párrafos 12, 17 y 39.

⁸ Artículo 10 del [Convenio Europeo de Derechos Humanos](#), 4 de septiembre de 1950; artículo 9 de la [Carta Africana sobre los Derechos Humanos y de los Pueblos](#) (Carta de Banjul), 27 de julio de 1981; artículo 13 de la [Convención Americana sobre Derechos Humanos](#), 22 de noviembre de 1969. Para una explicación del test tripartito, véase el siguiente video de YouTube: UNESCO, ‘[Los límites legítimos de la libertad de expresión: La prueba de tres partes](#)’.

⁹ El manual ‘[Discurso de Odio](#)’ de ARTICLE 19 ofrece una guía para identificar el ‘discurso del odio’ y cómo contrarrestarlo efectivamente al tiempo que se protegen los derechos a la libertad de expresión y a la igualdad. Véase también UNESCO, [Hacer frente al discurso de odio en las redes sociales: desafíos contemporáneos](#), y el video de YouTube de UNESCO ‘[¿Cómo abordar el discurso de odio en línea con un enfoque basado en los derechos humanos?](#)’.

¹⁰ Véase también el video de la UNESCO ‘[Los límites legítimos a la libertad de expresión: El Plan de Acción de Rabat](#)’.

¹¹ A lo largo de este informe nos referimos y citamos las normas comunitarias, los términos de servicio y otros documentos de plataformas de redes sociales como eran al momento de escribir. Se trata de documentos vivos que cambian y se actualizan constantemente, así que es posible que su contenido se haya modificado.

¹² Para conocer más sobre los distintos tipos de información falsa, véase UNESCO, [Periodismo, “noticias falsas” & desinformación: manual de educación y capacitación en periodismo](#).

¹³ El Consejo de Europa, en su informe [Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making](#) [Trastornos de la información: hacia un marco interdisciplinario para la investigación y la elaboración de políticas], propone la siguiente distinción entre 'desinformación', 'información errónea' e información falsa: 'desinformación' se refiere a la circulación de información falsa con el ánimo de causar daño; 'información errónea' se refiere a información falsa pero no se creó con el fin de causar daño; 'malinformación' se trata de información basada en la realidad que se usa con el fin de causar daño.

¹⁴ Ejemplos recientes incluyen [Turquía](#), [Túnez](#), [Sudán](#) y el [Reino Unido](#).

¹⁵ En este contexto, es importante señalar que es esencial crear un entorno propicio para los derechos a la libertad de expresión y a la igualdad que aborde las causas subyacentes tanto de la 'desinformación' como de los 'discursos de odio'. Los Estados deben, por ejemplo, centrarse en sus obligaciones positivas de promover un entorno de comunicaciones libre, independiente y diverso, incluyendo la diversidad de medios y la alfabetización digital y mediática, que son medidas clave para abordar la 'desinformación' y los 'discursos de odio'. Para más detalles, véase UNESCO, [Contrarrestar el discurso de odio](#); ARTICLE 19, [Manual sobre el Discurso de Odio](#); ARTICLE 19, [Submission to UN Special Rapporteur on Freedom of Expression and "Disinformation"](#) [Informe presentado ante el Relator Especial de la ONU sobre libertad de expresión y 'desinformación']; y UNESCO, [Ciudadanía alfabetizada en medios e información: pensar críticamente, hacer clic sabiamente](#).

¹⁶ [Principios Rectores sobre las Empresas y los Derechos Humanos: Puesta en práctica del marco de las Naciones Unidas para 'proteger, respetar y remediar'](#), elaborados por el Representante Especial del Secretario General para la cuestión de los derechos humanos y las empresas transnacionales y otras empresas comerciales. El Consejo de Derechos Humanos de la ONU suscribió los *Principios Rectores* en su [resolución 17/4 de 16 de junio de 2011](#).

¹⁷ [Principios Rectores de la ONU](#), Principio 15.

¹⁸ Consejo de Derechos Humanos de la ONU, [Informe del Relator Especial para la Libertad de Expresión](#), 6 de abril de 2018, A/HRC/38/35, párrs. 45-46.

¹⁹ Doc. ONU, [A/HRC/38/35](#), párr. 52.

²⁰ Doc. ONU, [A/HRC/38/35](#), párrs. 28 (para la proporcionalidad de las sanciones) y 59 (para la eficacia de los recursos).

²¹ En diciembre de 2022 el Tribunal de Roma falló a favor de Meta y revocó las decisiones de 2019 y 2020. En el fallo el Tribunal consideró que la libertad de expresión encuentra "como límite inmanente el respeto por los derechos fundamentales de los demás, en primer lugar la dignidad humana". Por tal razón, concluyó que Facebook además de estar facultada para terminar la relación contractual, estaba legalmente obligada a remover el contenido. Véase sentenza di primo grado n. 17909 del 5 diciembre 2022 del Tribunal de Roma.

²² Doc. ONU, [A/HRC/38/35](#), párr. 11.

²³ Doc. ONU, [A/HRC/38/35](#), párr. 50.

²⁴ Doc. ONU, [A/HRC/38/35](#), párr. 51.

²⁵ Doc. ONU, [A/HRC/38/35](#), párr. 52.

²⁶ Véase ARTICLE 19, [Vigilar a los vigilantes](#), p. 34.

²⁷ ARTICLE 19, [Content Moderation and Freedom of Expression](#) [Moderación de contenido y libertad de expresión], p. 18.

²⁸ Para una serie de buenas prácticas, véase Spandana Singh y Kevin Bankston, [The Transparency Reporting Toolkit: Content Takedown Reporting](#) [Caja de Herramientas sobre los Reportes de Transparencia: Reportes en materia de remoción de contenido].

²⁹ ARTICLE 19, [Side-Stepping Rights](#) [Eludiendo los derechos], pp. 16-17.

³⁰ ARTICLE 19, [Intermediarios de internet: Disyuntiva por la atribución de responsabilidad civil y penal](#), p. 3.

³¹ En primera instancia, la corte estableció que, como administrador, Chiranuch Premchaiporn era responsable de monitorear el contenido del foro en tanto podría afectar negativamente los derechos y libertades de terceros y la seguridad nacional. Las instancias superiores compartieron esta decisión.

³² Para un resumen del caso, véase Columbia University Global Freedom of Expression, [Prosecutor v. Chiranuch Premchaiporn](#).

³³ Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión y, [Informe del 16 de mayo de 2011](#), A/HRC/17/27, párr. 43.

³⁴ Doc. ONU, [A/HRC/17/27](#), para. 42.

³⁵ Véase, por ejemplo, la [Online Safety Bill](#) [Ley de Seguridad en Línea] del Reino Unido o la [Ley de Servicios Digitales](#) de la Unión Europea.

³⁶ Véase, por ejemplo, la [Online Safety Bill](#) [Ley de Seguridad en Línea] del Reino Unido (aunque la disposición sobre 'legal pero perjudicial' para adultos ha sido reemplazada con la obligación de que las empresas hagan cumplir sus términos de servicio) o el [Proyecto de Ley para Luchar contra el Contenido de Odio en Línea](#) (conocido como *Loi Avia*) que el Consejo de Estado francés declaró inconstitucional.

³⁷ J. Mchangama and J. Fiss, [The Digital Berlin Wall: How Germany \(Accidentally\) Created A Prototype for Global Online Censorship](#) [El muro de Berlín digital: Cómo Alemania creó (accidentalmente) un prototipo para la censura global en línea]; véase también '[The Digital Berlin Wall: How Germany \(Accidentally\) Created a Prototype for Global Online Censorship – Act Two](#)' [El muro de Berlín digital: Cómo Alemania creó (accidentalmente) un prototipo para la censura global en línea – Segundo Acto].

³⁸ ARTICLE 19, [Content Moderation and Local Stakeholders in Bosnia and Herzegovina](#) [Moderación de contenido y partes interesadas en Bosnia y Herzegovina], p. 39.

³⁹ Artículo 17 de la propuesta de [Ley Europea de Libertad de los Medios de Comunicación](#).

⁴⁰ Artículo 17 de la propuesta de [Ley Europea de Libertad de los Medios de Comunicación](#).

⁴¹ La comparación de *hash* asigna una 'huella digital' única a las imágenes y videos perjudiciales que se hayan detectado previamente. Los nuevos contenidos perjudiciales generados por los usuarios se pueden

eliminar automáticamente si el hash calculado coincide con un *hash* almacenado en la base de datos de contenidos perjudiciales conocidos. Véase Ofcom, [Use of AI in Online Content Moderation](#) [Uso de la IA en la moderación de contenidos en línea], p. 48.

⁴² Aunque las tecnologías de *hash* digital utilizan el reconocimiento de imágenes, la técnica también puede emplearse de forma más amplia, por ejemplo, para identificar objetos concretos dentro de una imagen, como un arma. Véase S. Singh (2019) '[Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content](#)' [Todo sobre la moderación: Un análisis de cómo las plataformas de internet usan Inteligencia Artificial para moderar el contenido generado por los usuarios], *New America*, July, p. 14.

⁴³ El Procesamiento de Lenguaje Natural es una técnica que consiste en analizar texto para hacer predicciones sobre su significado (por ejemplo, si expresa una opinión positiva o negativa). Véase Center for Democracy and Technology, '[Mixed Messages? The Limits of Automated Social Media Content Analysis](#)' [¿Mensajes contradictorios? Los límites del análisis automatizado del contenido de las redes sociales], p. 9.

⁴⁴ Singh, '[Everything in Moderation](#)' [Todo sobre la moderación], p. 7.

⁴⁵ Singh, '[Everything in Moderation](#)' [Todo sobre la moderación], p. 7.

⁴⁶ Center for Democracy and Technology, '[Mixed Messages?](#)' [¿Mensajes contradictorios?], p. 5.

⁴⁷ Singh, '[Everything in Moderation](#)' [Todo sobre la moderación], p. 13.

⁴⁸ Singh, '[Everything in Moderation](#)' [Todo sobre la moderación], p. 18.

⁴⁹ ARTICLE 19, [Content Moderation](#) [Moderación de contenidos], p. 17; Brennan Center for Justice, [Double Standards in Social Media Content Moderation](#) [Doble estándar en la moderación de contenidos en redes sociales], p. 18.

⁵⁰ Center for Democracy and Technology, '[Mixed Messages?](#)' [¿Mensajes contradictorios?], p. 6.

⁵¹ Center for Democracy and Technology, '[Mixed Messages?](#)' [¿Mensajes contradictorios?], p. 14.

⁵² Singh, '[Everything in Moderation](#)' [Todo sobre la moderación], p. 16.

⁵³ B. Heller (2019) '[Combating Terrorist-Related Content through AI and Information Sharing](#)' [Lucha contra los contenidos relacionados con el terrorismo mediante IA y el intercambio de información], Institute for Information Law, 26 de abril, p. 3.

⁵⁴ ARTICLE 19, [Side-Stepping Rights](#) [Eludiendo los derechos], p. 32.