derecom

ISSN: 1988-2629. Nº. 32. Nueva Época. Marzo 2022-Septiembre 2022

Cómo citar este texto:
Herrán Aguirre, A. F. y P. Pazos, M. H. (2022). Regulación de redes sociales para evitar la violencia contra las mujeres, *Derecom*, 32, 89-112, http://www.derecom.com/derecom/

REGULACIÓN DE REDES SOCIALES PARA EVITAR LA VIOLENCIA CONTRA LAS MUJERES

REGULATING SOCIAL MEDIA TO PREVENT VIOLENCE AGAINST WOMEN FROM HAPPENING

© Alejandro F. Herrán Aguirre
Universidad Autónoma de Chiapas (México)
alejandro.herran@iij-unach.mx

©Maricela Hazel P. Pazos
Universidad Autónoma de Chiapas (México)
haz.pazos@iij-unach.mx

Resumen

El auge que las redes sociales han tenido en los últimos años hace un llamado a reflexionar sobre la necesidad de regular el contenido que se comparte en ellas. Las mujeres, especialmente las niñas y adolescentes, padecen mayores experiencias de violencia digital a través de las redes sociales. En este trabajo se hace un repaso de la violencia digital de la que las mujeres son víctimas y sus efectos —tanto individualmente como para la sociedad, e incluso de la posibilidad de violencia en la vida real—. Se hace también un recuento de las principales estrategias que se han implementado para regular las redes sociales, usando la perspectiva de género como herramienta de análisis. La moderación de contenido ya sea hecha por algoritmos o seres humanos es una tarea necesaria y difícil, a través del estudio de sus componentes y sus dificultades el trabajo permite comprender los problemas de la moderación —especialmente a la luz de los principios de la libertad de expresión— y los retos para la regulación, la atribución de responsabilidades y, más importante aún, para la protección de los grupos vulnerables como las mujeres.

Recibido: 15/02/2022 Revisado: 03/03/2022
Aceptado: 14/03/2022 Publicado:15/03/2022

Summary

The boom that social networks have had in recent years calls for reflection on the need to regulate the content that is shared on them. Women, especially girls and teenagers, suffer harder experiences of digital violence through social networks. In this paper we review the digital violence of which women are victims and its effects —both individually and for society, and even the possibility of violence in real life—. A review is also made of the main strategies that have been implemented to regulate social networks, using the gender perspective as an analysis tool. Content moderation, whether done by algorithms or human beings, is a necessary and difficult task. Going through its components and its difficulties, the paper makes it easier to understand the issues of moderation —especially considering the principles of freedom of expression— and the challenges for regulation, attribution of responsibilities and, more importantly, for the protection of vulnerable groups such as women.

Palabras clave: Redes Sociales. Regulación. Perspectiva de Género. Violencia Digital.

Keywords: Social Media. Regulation. Gender Perspective. Digital Violence.

1.Introducción

Las redes sociales han crecido en importancia conforme su uso se ha generalizado. Su influencia en la vida de las personas e instituciones puede ser poderosa. Desde los años 90 del siglo pasado, cuando el internet comenzaba a desplegarse como un factor transformador de la comunicación, personas de todo tipo auguraron el surgimiento de poderosas herramientas que permitirían la comunicación directa entre todos. Se prometía un futuro donde la facultad de conexión entre personas multiplicaría los beneficios sociales, fortalecería las relaciones personales e incluso reforzaría la democracia. El tiempo ha mostrado que, aunque muchas de las predicciones positivas se han cumplido, el internet ha creado nuevos espacios de conflicto donde lo peor de la interacción humana puede ser potenciado. Es en este clima de dificultad donde se llama a la regulación de las redes sociales como plataformas de comunicación. En este trabajo presentamos una parte importante del desafío de la moderación de contenido y de la regulación de las redes sociales, aquella relacionada con un grupo en situación de vulnerabilidad: las mujeres.

En el primer apartado se analiza de manera breve la violencia digital contra las mujeres, presentando datos que ponen en evidencia que las mujeres viven de manera diferenciada la violencia en el espacio virtual y en las redes sociales en particular, y algunas de estas manifestaciones. Después se presentan dos principales versiones de moderación, la autoimpuesta y la externa, mostrando la dificultad técnica en sus implementaciones y su relación con los incentivos que existen para su ejecución. También se comentan las principales estrategias de regulación implementadas en el mundo, como el caso del producto legislativo de Alemania. Finalmente, se hacen algunas reflexiones sobre lo que se necesita para una regulación

con perspectiva de género, conscientes de que es un tema complejo, en el que no se pueden ofrecer soluciones sencillas, porque no las hay.

2.Violencia en Línea contra las mujeres

El espacio virtual, y con ello las redes sociales, son un reflejo, o incluso, una extensión del espacio tradicional. Actualmente es difícil imaginar la vida sin las redes sociales, especialmente a partir del confinamiento ocasionado por la pandemia del SARS-COV-2, que llevó a gran cantidad de actividades a trasladarse al espacio virtual.

Entender a las redes sociales como una extensión de nuestra vida implica reconocer que reproducen las situaciones tanto buenas como malas. En ese sentido, también reproducen la discriminación y violencia que viven las mujeres (Harris y Vitis, 2020). Sin embargo, la virtualidad presenta desafíos propios, especialmente a nivel jurídico, debido a las características que posee. El entorno virtual tiene la capacidad de maximizar esta violencia y discriminación.

Aunque la violencia y discriminación en redes sociales puede vivirla cualquier persona, son las que pertenecen a grupos en situación de vulnerabilidad quienes la afrontan con mayor frecuencia —igual que en el mundo tradicional—.

La evidencia empírica en diversas áreas relacionadas con la discriminación y la violencia, y su relación con las redes sociales ha crecido. En relación con la expresión y la discriminación se ha mostrado que las redes sociales pueden contener expresiones discriminatorias encubiertas. Ben-David y Matamoros-Fernández analizaron expresiones publicadas en Facebook por siete partidos políticos de extrema derecha en España y concluyeron que *Facebook aloja un creciente volumen de prácticas discriminatorias encubiertas que no solo circulan datos y contenido, sino que también denotan discurso de odio no encubierto por parte de los seguidores de las páginas* (Ben-David y Matamoros-Fernández, 2016, p. 1188). De manera relacionada se ha vinculado la prevalencia de crímenes de odio con el uso de redes sociales (Relia y otros, 2019). Así también se ha mostrado que, aunque el contenido de odio puede representar un porcentaje bajo del total de expresiones estudiadas, es más visible que otras formas de contenido (Kaakinen y otros, 2017).

Los adolescentes son otro grupo importante por su relación con las redes sociales. Flores y otros (2017) documentaron el peligro de su invasión a la privacidad. Así también se ha mostrado que la influencia negativa de las redes sociales en su vida puede conducir a afectaciones en patrones de sueño, desórdenes alimenticios, hostigamiento, consumo de alcohol, etc. (Villanueva y otros, 2017).

Respecto de las mujeres como grupo vulnerable la Comisión de la Banda Ancha para el Desarrollo Digital de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, en un reporte de 2015 presenta datos que indican que las personas usuarias con nombres femeninos sufren más amenazas que quienes tienen nombres masculinos y que son las personas entre 18 y 24 años las más afectadas (UNESCO, 2017, p. 15). Amnistía Internacional, a través de una encuesta realizada por IPSOS MORI, estudió las experiencias de mujeres en edades de entre 18 y 55 años en 8 países (Dinamarca, Italia, Nueva Zelanda, Polonia, España, Suecia, el Reino Unido y los Estados Unidos de América). El 23% de las mujeres entrevistadas dijo haber experimentado abuso o acoso en línea al menos una vez. El 46% dijo que la naturaleza del abuso o acoso que sufrieron en línea fue sexista y 41% dijo que la experiencia les hizo sentir

que su seguridad física estaba en peligro (Amnistía Internacional, 2017). En materia deportiva, las redes sociales promueven el acoso y odio basado en el género hacia las mujeres en maneras que los medios tradicionales no lo hacen (Kavanagh y otros, 2019). De manera similar Woodlock y otros (2019) mostraron el importante papel que la tecnología tiene en la perpetración de la violencia doméstica contra las mujeres.

Tal vez más preocupante sea el impacto negativo de las redes sociales en las niñas y las adolescentes. Estudios recientes arrojan evidencia que indica la correlación entre el uso de redes sociales y daños psicológicos como depresión, desórdenes alimenticios, y otros, afectando principalmente a niñas y adolescentes (Haidt y Twenge, 2020; Twenge y otros, 2022). En coincidencia, durante el año 2021, Frances Haugen —cuya historia se detalla más adelante—, entregó a algunos periodistas documentos internos de la plataforma que indicaban que Facebook ignoró los resultados de investigaciones propias que habían encontrado que Instagram, como plataforma, tenía efectos negativos en la salud mental de niñas y adolescentes (Wells y otros, 2021). Adicionalmente, Plan International publica un informe sobre las experiencias en línea y redes sociales de niñas de todo el mundo. En su tercer reporte entrevistaron a 14,071 niñas y mujeres jóvenes de entre 15 y 25 años de edad de 22 países. Los hallazgos principales indican que más de la mitad de la niñas entrevistadas ha sufrido abuso o acoso en línea. Como resultado de ello 1 de cada 4 niñas víctimas de abuso en línea teme por su seguridad física. En consonancia con los resultados de investigaciones realizadas en materia de expresión, el abuso tiene un efecto acallador, silenciando las voces de niñas en el espacio virtual (Plan International, 2020).

Los trabajos citados, que forman parte de un cuerpo creciente de investigación, muestran que los efectos negativos de las redes sociales en las mujeres y otros grupos vulnerables deben ser tomados en serio, no solo por investigadores, sino por todos los actores políticos y jurídicos relacionados con el problema.

Desde una perspectiva positiva las redes sociales pueden ser herramientas para la promoción y disfrute de los derechos humanos. Sin embargo,

> "*existe un riesgo considerable de que el uso de las TIC sin aplicar un enfoque basado en los derechos humanos y la prohibición de la violencia en línea por razón de género puedan llevar a un aumento aun mayor de la discriminación sexual y por razón de género, y de la violencia contra las mujeres y las niñas en la sociedad*" (ONU 2018).

Por lo que aún la promoción de las virtudes de la red debe hacerse considerando los efectos en la vida real —incluyendo los negativos— que amplifican.

Esta realidad ha despertado la preocupación de instituciones internacionales como la Comisión Interamericana de Derechos Humanos (CIDH), que señala que *la violencia contra las mujeres en internet ha surgido como una nueva forma de violencia por razones de género la cual la CIDH nota se está extendiendo rápidamente y supone un peligro significativo*." (CIDH, 2019)

El incremento de estos ataques y sus características específicas requiere hablar de manera particular de violencia en línea contra las mujeres, entendida como

> *todo acto de violencia por razón de género contra la mujer cometido, con la asistencia, en parte o en su totalidad, del uso*

> *de las TIC, o agravado por este, como los teléfonos móviles y los teléfonos inteligentes, Internet, plataformas de medios sociales o correo electrónico, dirigida contra una mujer porque es mujer o que la afecta en forma desproporcionada. (ONU, 2018)*

En este sentido entonces, es claro que cuando nos referimos a estas conductas *"aludimos a las disímiles formas en que, a través de las TICs, se pretende causar o se causa daño a una o varias personas, con el objetivo de preservar o reforzar la dominación masculina y mantener la subordinación femenina mediante los pactos patriarcales en pos de la desigualdad y la opresión de género* (García Román y Mindek Jagic 2021).

Las consecuencias que esta violencia provoca en las personas pueden ser muy variadas, pues atienden a factores como el tipo de conducta realizada, si la persona que realiza la agresión es conocida o no por la víctima, la actividad que las víctimas realizan en las redes sociales, la importancia y el tiempo que le dedican a estos espacios, entre otras; sin embargo, hay algunos efectos que han sido estudiados por diversas organizaciones, entre ellas, la Organización de las Naciones Unidas, que señala que algunas de las consecuencias individuales son: abstenerse de usar Internet, aislamiento social —que lleva a las víctimas o supervivientes a retirarse de la vida pública, incluso la familia y las amistades—, y la movilidad limitada, es decir, la pérdida de libertad para desplazarse en condiciones de seguridad. También pueden presentarse consecuencias psicológicas, como: niveles más altos de ansiedad, trastornos de estrés, depresión, trauma, ataques de pánico, pérdida de autoestima y una sensación de impotencia en su capacidad para responder al abuso. (ONU, 2018)

Además de los efectos individuales que puede causar esta violencia también genera *una sociedad en que las mujeres ya no se sienten seguras en línea o fuera de línea, debido a la impunidad generalizada de los autores de la violencia de género, p*ues *el acoso en internet, especialmente mediante amenazas en plataformas de redes sociales, se ha consolidado en el periodo como una forma para intimidar, infundir miedo y censurar* (ONU Mujeres 2020).

3.Tipos de violencia digital contra las mujeres

Un aspecto que diferencia esta violencia contra las mujeres es que los ataques se caracterizan por utilizar estereotipos sexistas, descalificación y amenazas hacia la familia y las relaciones personales, comentarios negativos sobre la apariencia física y agresiones sexuales.

A partir de lo anterior, diversas organizaciones, autores y autoras han hecho clasificaciones de los tipos de violencia que las mujeres viven en las redes. Esta clasificación puede ser extensa e incluir muchas conductas. Aunque no es el objetivo en este trabajo hacer una lista exhaustiva, es apropiado señalar algunas de las clasificaciones sobresalientes. Como primer punto es importante notar que los avances en la tecnología con frecuencia resultan en que las clasificaciones se modifiquen rápidamente; por lo que lo esencial es comprender que estas agresiones vulneran diversos derechos de las personas usuarias tales como el derecho a una vida libre de violencia, el derecho a la no discriminación, el derecho a la libertad, el derecho a la privacidad y a la protección de datos.

La Organización de Estados Americanos publicó en 2021 el manual *La violencia de género en línea contra las mujeres y niñas. Guías de conceptos básicos, herramientas de seguridad digital y estrategias de respuesta* (OEA, 2021). Es un trabajo general sobre el tema

que incluye un recorrido por los tipos de violencia de género contra las mujeres y las niñas facilitada por las nuevas tecnologías. La lista incluye las siguientes conductas: 1) Creación, difusión, distribución o intercambio digital de fotografías, videos o audioclips de naturaleza sexual o íntima sin consentimiento; 2) Acceso, uso, manipulación o distribución no autorizadas de datos personales; 3) Suplantación y robo de identidad; 4) Actos que dañan el honor o la credibilidad de una persona; 5) Actos que implican la vigilancia y el monitoreo de una persona; 6) Ciberhostigamiento; 7) Ciberacoso; 8) Ciberintimidación; 9) Amenazas directas de daño; 10) Violencia física facilitada por las tecnologías; 11) Abuso, explotación de mujeres y niñas a través de las tecnologías; y, 12) Ataques a grupos, organizaciones o comunidades de mujeres.

En la doctrina, es importante la aportación realizada por García Román y Jagic (2021), quienes presentan la siguiente lista: a) Ciberacoso (*stalking,* en inglés); b) *Ciberbullying* (ciberacoso entre pares); c) *Grooming* (acoso y abuso sexual *online* contra menores); d) Cibercontrol; e) *Doxing*; f) Hate Speech (discurso de odio); g) *Flaming*; h) *Gossip* (chisme en inglés); i) Difusión de contenido íntimo a través de medios tecnológicos sin consentimiento; j) Suplantación de identidad; y, k) *Slut-Shaming* (tildar de "mala mujer"). Por otra parte, la Comisión de Derechos Humanos de la Ciudad de México (2021) enlista los siguientes tipos de violencia contra las mujeres en el espacio digital: daños a la integridad física; daños a la integridad psicoemocional y daños a la esfera social, profesional y económica. Como es evidente, y como se dijo anteriormente, existen muchas clasificaciones diferentes de la violencia digital contra mujeres. En las mostradas en este trabajo se puede apreciar que comparten elementos comunes, aunque también es posible apreciar diferencias sustanciales. El estudio de las clasificaciones y conductas de violencia digital contra mujeres es constante y, como es evidente, un análisis detallado de las clasificaciones excedería el espacio de este trabajo.

A continuación, se describirán algunas de las agresiones que se cometen más frecuentemente contra las mujeres con el propósito de que comprendamos las implicaciones de género que pueden tener estas conductas en las redes sociales. Las agresiones que a continuación se comentan no representan la totalidad de las que son estudiadas por la doctrina. Como es evidente, la variedad de diferencias en las clasificaciones imposibilita la definición y estudio en este trabajo. Las agresiones que se describen a continuación representan algunas de las manifestaciones de violencia digital más importantes en redes sociales y con frecuencia son citadas como fundamento para la regulación de estas.

4.Agresiones contra la privacidad.

Aunque todas las personas pueden sufrir ataques a su privacidad en las redes sociales, es un hecho que las mujeres resultan particularmente afectadas.[1]

Esta agresión implica compartir o publicar, sin la autorización de la persona propietaria, cualquier tipo de información o datos. En este sentido, la información que se comparte puede ser sensible como son los datos personales, domicilios, nombres de familiares, ingresos económicos o cualquier documento que sea privado. Dicha acción puede poner en riesgo a la persona que ha sido víctima. La difusión de información sin consentimiento puede tener diversas manifestaciones. Las siguientes son algunas de ellas: *doxing,*[2] robo de contraseñas, utilización de programas espías, intervención/escucha en sus dispositivos, robo de equipo, bloqueo de acceso propio, *phishing,*[3] suplantacion y robo de identidad.

5.Agresiones sexuales.

Como señalamos al principio, una de las principales manifestaciones de violencia contra las mujeres en redes sociales tiene que ver con contenido sexual, tanto expresiones y amenazas de realizar ataques sexuales, como difusión de contenido íntimo —imágenes, audios o videos— sin autorización. Es importante distinguir la actividad de compartir imágenes íntimas con el consentimiento de las personas involucradas (*sexting),*[4] con la difusión no consentida. En el segundo caso la conducta reprochable consiste en la diseminación de contenido íntimo privado. El contenido íntimo puede haber sido obtenido con el consentimiento de la víctima o sin ella. A este tipo de conducta se la ha llamado "pornovenganza". Sin embargo, el término no es apropiado porque no describe el daño que se realiza a la víctima, entre otras razones. La importancia y severidad de esta conducta ha llevado a su determinación como delito, la forma y particularidades de la tipificación —así como de las penas— varía entre las diversas jurisdicciones en que se ha implementado. Por ejemplo, en México, se realizaron diversas reformas legislativas con el propósito de reconocer la violencia digital y sancionar los delitos relacionados con ella.[5] Las reformas, al momento de redacción, se han realizado al Código Penal Federal y a los Códigos Penales de 29 entidades federativas. En el artículo 199 octies del Código Penal Federal se tipifica el delito de Violación a la Intimidad Sexual, indicando que lo comete *aquella persona que divulgue, comparta, distribuya o publique imágenes, videos o audios de contenido íntimo sexual de una persona que tenga la mayoría de edad, sin su consentimiento, su aprobación o su autorización*. En los Estados Unidos, son varios los Estados que han tipificado esta conducta como delito, aunque con variaciones sustanciales (Franks, 2015). En España, es en el artículo 197.7 del Código Penal donde se encuentra tipificada esta conducta. en él se castiga a quien *sin autorización de la persona afectada, difunda, revele o ceda a terceros imágenes o grabaciones audiovisuales de aquélla que hubiera obtenido con su anuencia […], cuando la divulgación menoscabe gravemente la intimidad personal de esa persona*.

Aunque estas conductas pueden realizarse contra cualquier persona, es especialmente agresiva contra las mujeres, pues la forma en la que hombres y mujeres expresan su sexualidad es diferente. Mientras que socialmente es aceptado que los hombres tengan muchas parejas sexuales y que incluso, presuman de ello, a las mujeres se les juzga por ejercer su sexualidad libremente.

Además, cuando una mujer es víctima de esta forma de violencia digital, son vistas como objetos, y se les agrede y culpabiliza de la agresión recibida, simplemente por estar ahí, o por estar de esa manera, o por usar esa red social, o sencillamente por ser mujeres. Algunos estudios revelan que el 90% de las víctimas de la distribución digital no consensuada de imágenes íntimas son mujeres (ONU, 2018).

Otra forma de violencia digital relacionada con la sexualidad contra las mujeres es conocida como d*eepfake*, término que surgió en 2017, para referirse a videos que utilizan técnicas de aprendizaje automático para intercambiar la cara de una persona con la de otra. La cantidad de *deepfakes* en línea está creciendo exponencialmente y se debe en parte al hecho de que ahora es más fácil para los no expertos usar ciertas tecnologías.

Aunque se trata de tecnología que puede utilizarse con cualquier fin, y contra cualquier persona, destaca su creciente uso en contextos políticos y pornografía. Los *deepfakes* sexuales son una de las muchas formas en que las mujeres son cosificadas en la cultura digital y visual.

De acuerdo con un informe de Deeptrace, en cuanto a su uso en pornografía, son las mujeres las víctimas exclusivas (Deeptrace 2019).

6.Agresiones discriminatorias y discurso de odio.

Se trata de un tipo de discurso que refleja patrones culturales que asignan un rol secundario o únicamente reproductivo (y/o sexual/sexualizado) a las mujeres y a otros cuerpos. Pueden o no incitar a la violencia. Es una forma de violencia simbólica basada en las ideas preconcebidas tradicionales de género. También es prevalente el discurso de odio, que puede ser definido como las expresiones destinadas a intimidar, oprimir, incitar al odio o la violencia contra una persona o un grupo con base en la raza, religión, nacionalidad, género, orientación sexual, discapacidad o alguna otra característica (Esquivel, 2016). La discriminación y el discurso de odio no son exclusivos de internet pero factores como la expansión de la tecnología así como la posibilidad de transmitir mensajes fácilmente han contribuido a que grupos que antes existían en fragmentos separados puedan conectarse y generar un sentido de comunidad (Hernández, 2020). El discurso de odio puede tener diversas manifestaciones en el espacio virtual. Puede incluir expresiones orales, escritas, imágenes o videos (Campos y otros, 2015).

7.Libertad de expresión en el contexto internacional

La violencia en línea contra las mujeres, por el medio en que se realiza, es una forma de expresión. La intervención necesaria para detener o mitigar los efectos negativos de la expresión en redes sociales puede darse, en la perspectiva amplia, por dos caminos: la moderación del contenido, es decir, las actividades que hacen las plataformas de redes de forma rutinaria para determinar qué contenidos deben eliminarse; y la regulación, que implica un esfuerzo legislativo o de otro tipo de control externo sobre las empresas de redes sociales para influir en el contenido que circula por ellas. En ambos casos el principal problema es determinar cómo debe definirse el contenido que debe eliminarse y este es un tema de libertad de expresión.

En el ámbito internacional la libertad de expresión sigue principios generales que la protegen como derecho fundamental, tanto para la autodeterminación de la persona como para la construcción de la sociedad democrática. El artículo 19 del Pacto Internacional de Derechos Civiles y Políticos, que ha sido ratificado por 173 países, contiene las disposiciones generales respecto de la libertad de expresión e incluye las causas justificadas para restringirla: asegurar el respeto a los derechos o a la reputación de los demás, y la protección de la seguridad nacional, el orden público, la salud o la moral pública; indicando también que toda restricción deberá ser fijada expresamente por la ley. De manera similar, el artículo 13 de la Convención Americana de Derechos Humanos contiene la protección de la expresión, entendida también como el derecho a la información y contiene una formulación similar de posibles restricciones. En otras palabras, el régimen de limitación de la expresión requiere que los países construyan las limitaciones sobre dos requisitos esenciales. Primero, las limitaciones a la expresión deben estar contenidas en la ley, y, tal vez más importante, las limitaciones deben perseguir alguno de los fines determinados por el Derecho Internacional, estos incluyen la protección de derechos de otras personas, de la seguridad nacional, del orden público, etc. La libertad de expresión, por su importancia para el desarrollo individual y para la sociedad democrática, existe bajo un régimen de excepción y bajo la prohibición de censura previa, lo que implica que la responsabilidad que se impute por expresiones debe ser siempre ulterior; el objetivo debe ser limitar la expresión lo menos posible: solo en los casos en los que no hacerlo producirá un daño y las responsabilidades deben ser siempre ulteriores. Entonces, para que se pueda dirigir a las plataformas de redes sociales hacia la eliminación de cierto tipo de contenido la regulación debe cumplir los requisitos de los fundamentos generales de la libertad de expresión.

De acuerdo con lo anterior, es evidente que la limitación de la expresión —o la imposición de responsabilidades— en aras de la integridad física y psicológica de personas en condición de vulnerabilidad cumple con los requisitos del Derecho Internacional, por lo que la pregunta importante es ¿cómo puede protegerse a las personas de las expresiones dañinas que circulan por redes sociales? Para ofrecer una respuesta es necesario comprender cuáles son los mecanismos actuales mediante los que se atribuye responsabilidad a las personas que hacen expresiones no protegidas.

En relación con la violencia en redes sociales, se ha documentado que existen varios tipos de expresiones que pueden ser consideradas como tales de acuerdo con su contenido. Estas son las expresiones discriminatorias, los insultos gratuitos y, en general, aquellas que pueden ser descritas como discurso de odio (Malmasi y otros, 2017).[6] Regular o moderar estas expresiones en redes sociales requiere construir criterios de limitación que no contravengan los fundamentos de la libertad de expresión. Esto es complicado.

En términos generales, es posible decir que existen dos grandes visiones sobre la limitación de la expresión en cuanto toca a expresiones de odio, la de los Estados Unidos de América —bajo la Primera Enmienda a su Constitución— y la de Europa. En general, la visión de la Primera Enmienda admite muy pocas limitaciones a la expresión y no reconoce al discurso de odio como categoría de expresión limitada. La jurisprudencia de Estados Unidos sobre la libertad de expresión es basta y complicada, pero es posible afirmar que de forma general el estudio de la constitucionalidad de las expresiones, es decir, cómo delimitar qué expresiones están protegidas por la constitución y cuáles no, ha seguido dos tipos de análisis. El objetivo del análisis constitucional es determinar si alguna limitación a la expresión —usualmente de tipo legal— viola o no la Primera Enmienda de su Constitución. Así que las limitaciones son analizadas dependiendo de si la limitación se da directamente al contenido de la expresión —se limita lo que se dice—, o si la limitación es acerca de algo diferente del contenido —se limita el lugar, el tiempo, la forma, etc.—. Las primeras limitaciones, las que estudian el contenido de la expresión, son analizadas mediante un test de escrutinio estricto muy difícil de superar y las segundas, que pueden ser llamadas limitaciones de tiempo, modo o lugar, son analizadas con un criterio menos exigente (Stephan, 1982). La Corte Suprema de los Estados Unidos ha sido exigente en la acreditación de los requisitos para el test de escrutinio estricto, por lo que las restricciones analizadas con este criterio son frecuentemente declaradas inconstitucionales. Esto ha resultado en una jurisprudencia muy permisiva de la expresión que permite pocas limitaciones basadas en el contenido.[7]

Por su parte, en Europa, se ha desarrollado un rico y complicado régimen de limitación de la expresión basado en el odio y la discriminación como lacerantes de la dignidad humana. El Tribunal Europeo de Derechos Humanos (TEDH) ha producido jurisprudencia acerca del discurso de odio y sus implicaciones para la sociedad (Bleich y Al-Mateen, 2021). En contrapunto, la jurisprudencia de los Estados Unidos se ha negado a reconocer el discurso de odio como una categoría para la limitación de la expresión. El precedente vigente en ese país es el expresado en Brandenburg vs. Ohio[8] (Heredia, 2017), en el cual se estableció la limitación de las expresiones que incitan a conductas ilegales directas e inminentes.

Es en este contexto internacional donde se desarrolla la expresión a través de las redes sociales. Las plataformas son actores privados que tienen actividad en muchos países. La regulación de las expresiones hechas en sus aplicaciones está sujeta al complicado sistema de

Derecho Internacional y nacional que, como se ha visto, tiene además diferentes visiones sobre la expresión y sus limitaciones.

8.Expresiones en redes sociales y su limitación

El problema de la regulación de las expresiones en redes sociales es determinar cómo puede identificarse la expresión problemática y, además, justificar su limitación sin contravenir los fundamentos de la libertad de expresión. Las plataformas de redes sociales derivan su valor principal de las publicaciones de las personas usuarias —esta es una de sus características definitorias (Carr y Hayes, 2015)— por lo que promueven la publicación del mayor tipo de contenidos. Las publicaciones de las personas usuarias pueden contener material que caiga en las restricciones a la libertad de expresión, ya sea por la ley o la jurisprudencia, o por los términos de uso (condiciones generales de contratación, en España) y las reglas específicas de la plataforma. La respuesta de las plataformas demanda la implementación de mecanismos de gobernanza que estructuren la participación de la comunidad para facilitar la cooperación y para prevenir el abuso. Es a esta implementación a la que se le denomina "moderación de contenido" (Grimmelman, 2015).

La moderación del contenido puede ser realizada por seres humanos que analicen las publicaciones individualmente —lo que permite mayor profundidad en el análisis—, pero también puede hacerse de forma automatizada. La automatización consiste en la implementación de sistemas que clasifican el contenido generado por las personas usuarias con base en el apareo[9] o la predicción, lo que conduce a un resultado determinado por una decisión de gobierno (Gorwa y otros, 2020, p. 3).[10] La implementación de los sistemas automatizados —comúnmente llamados algoritmos— en la moderación de contenidos es inevitable. La enorme cantidad de contenido generado por personas usuarias impide la moderación hecha únicamente por humanos. Los avances en la ciencia de la computación y el desarrollo de inteligencia artificial han generado amplias expectativas de la efectividad de los algoritmos como herramientas de moderación.

Sin embargo, existen importantes limitaciones y problemas en la implementación de algoritmos. Algunos de los problemas técnicos son evidentes; la identificación y apareo de contenido, por ejemplo, depende de la determinación de la base de datos que se usará en la moderación. Un cambio mínimo en el contenido que lo diferencie de la entrada en la base de datos producirá una identificación incorrecta. Este tipo de dificultades está siendo atendida en un activo campo de diseño de herramientas automatizadas y aunque constantemente se refinan los procedimientos de los algoritmos es imposible diseñar un sistema de identificación infalible (Grimmelman, 2020). Otra importante consecuencia negativa de la implementación de los sistemas automatizados consiste en la sobremoderación o en la moderación excesiva. Esto sucede cuando contenido lícito es afectado por el algoritmo por su cercanía o parecido con contenido prohibido. Esta dificultad complica la labor que desarrollan las plataformas, ya que deben hacer constantes alteraciones pequeñas tanto al contenido de la base de datos como a las reglas de operación del algoritmo. Como punto añadido no debe dejarse de lado el factor económico como incentivo en el diseño de las políticas de moderación. Si la plataforma es responsable del contenido o del error en la moderación, su incentivo será eliminar el contenido —en caso de duda, bloqueo— (Keller, 2015).

Además de los problemas técnicos es importante resaltar los humanos. Los algoritmos contienen en su diseño los sesgos de sus desarrolladores. Durante el diseño del algoritmo la persona o personas que desarrollen las reglas con las que se comportarán imprimirán sus heurísticas, visiones y otras subjetividades. Las personas llevan sus esquemas de pensamiento

personal y al procesar la información sesgan los resultados. Estos sesgos son más pronunciados en algunas categorías como el discurso de odio, contenido terrorista, o el *bullying* (Mackenzie, 2020). Esta consideración requiere que admitamos que los algoritmos —y por extensión, el proceso de moderación— no pueden ser imparciales ni objetivos.

Como se puede apreciar, los algoritmos están lejos de ser una solución única al problema de la moderación de contenido. Además de los problemas ya descritos tienen dificultad en identificar expresiones (Duarte y otros, 2018), problema que se ve complicado por el uso intencionado de palabras en clave para ocultar el significado de la expresión. La dificultad de la implementación resulta en que cuando es posible —si los recursos humanos, técnicos y económicos de la plataforma lo permiten—, el sistema de moderación de contenido usa tanto algoritmos como personas. Los primeros suelen ser la primera implementación de la moderación por su habilidad para procesar grandes cantidades de contenido. Los problemas o quejas que surgen del primer nivel son evaluados en un nivel superior por personas dedicadas a la moderación de contenido. Este modelo ha sido ya implementado en las redes sociales más importantes (Klonick, 2017).

En términos prácticos la moderación requiere que las redes sociales determinen una definición de expresiones no permisibles — principalmente, discurso de odio y expresiones discriminatorias—. Este es el punto de partida para los sistemas de moderación. Pero es posible ver que la definición que se use presentará problemas, ya sea por demasiada amplitud y aplicación a expresiones permitidas —lo que acallaría la expresión de millones de personas— , ya sea por ser demasiado restrictiva —en este caso correría el riesgo de no contemplar expresiones que sí merecen ser limitadas—. Este problema ha sido adecuadamente descrito por Siegel (2020) y representa uno de los puntos críticos en la discusión sobre la moderación de contenido en redes sociales. Siegel examinó la legislación de varios países para comparar distintas definiciones de discurso de odio e identificó algunos problemas de implementación. Por ejemplo, en el Reino Unido es un delito proferir ofensas que inciten al odio racial o religioso, pero en los Estados Unidos ese delito podría ser inconstitucional (p. 58). De manera similar una definición amplia, como la de Canadá, que caracteriza al discurso de odio como aquel que *intencionalmente promueve el odio contra cualquier grupo identificable* es muy difícil de implementar. Su uso por una plataforma de red social para moderar contenido indudablemente resultaría en la sobrerremoción de expresiones. Las definiciones amplias no solo son difíciles de implementar, como ya se mostró, sino que el régimen de responsabilidad y los incentivos económicos llevarían a su ejercicio dominante, los efectos serían el acallamiento de la expresión y la destrucción de los principales beneficios de las redes sociales. En el sentido opuesto, una definición que es demasiado restrictiva, como las que requieren intención de daño —o en el caso de Estados Unidos incitación a acciones ilegales inminentes— corren el riesgo de ser inútiles.

Además de lo anterior también es importante considerar la intención de las expresiones. Este punto es particularmente relevante para la técnica legislativa. Hay que distinguir entre aquellas expresiones hechas con el propósito de lastimar, provocar, incitar o promover violencia, de las que pueden ser vejatorias, pero constituyen solamente expresiones problemáticas. Esta distinción puede parecer difícil de hacer o “resbaladiza” pero la dificultad exhibe la naturaleza de los problemas de limitación de la expresión. La importancia de la distinción la expresa Sellars citado por Siegel (2020): *Expresiones que incitan a la violencia son diferentes de las que solo son “meramente” ofensivas, y el uso de lenguaje dañino por un único atacante es muy diferente de campañas coordinadas de odio de una muchedumbre digital*. La discusión de la distinción es importante porque permitirá una técnica legislativa más clara y directa. Los avances en la

automatización de la moderación son prometedores. Algunos investigadores han tenido éxito en diferenciar tipos de expresiones dirigidas a individuos o a grupos (Elsherief y otros, 2015).

Para definir qué tipo de expresiones deben ser limitadas para evitar violencia contra grupos en situación de vulnerabilidad es necesario usar las categorías sospechosas que la doctrina y la jurisprudencia han desarrollado al respecto. Estas incluyen la raza, el sexo, la orientación y expresión de género, el credo, la nacionalidad, entre otras,[11] lo cual permite identificar a las mujeres como uno de estos grupos que han sido históricamente discriminados.

Con frecuencia, el problema no radica en la determinación legal que prohíbe algún tipo de expresión, sino en la falta de aplicación de responsabilidades. Por ejemplo, países como Australia, Dinamarca, Francia, Alemania, India, entre otros, tienen leyes que prohíben diversos tipos de expresiones. Incluso tipifican como delito la incitación al odio racial o religioso, pero con respecto de internet estas normas no son aplicadas coherentemente.

Suprimir el contenido dañino o las cuentas de las personas que las emitieron es una tarea difícil. A veces, cuando las expresiones son cuantiosas y dañinas —y es posible identificar a la persona usuaria que las produce—, suprimir su cuenta o limitar su acceso a la plataforma parecen soluciones apropiadas. No obstante, como lo documenta Siegel, eliminar o bloquear usuarios puede ser inútil, ya que la persona, o personas, pueden crear nuevos perfiles o cuentas y continuar con la actividad. Este proceso puede resultar en la inversión de considerables recursos por parte de la plataforma en la identificación y eliminación de usuarios sin resultados efectivos. Es importante notar que el bloqueo de usuarios sí puede reducir la prevalencia de contenido problemático (Chandrasekharan y otros, 2017). Sin embargo, no es posible afirmar con convicción que los resultados sean permanentes. Es posible que los usuarios migren de plataforma, cambiando de foros públicos —como 4chan— a Facebook o Youtube; el seudoanonimato que ofrecen las plataformas protege a los usuarios problemáticos y dificulta la supresión o moderación de contenido dañino.

Este fenómeno es frecuente cuando se trata de ataques y agresiones contra las mujeres —el 73% de las mujeres en el mundo han estado expuestas o han experimentado algún tipo de violencia en línea (UNESCO, 2015, p. 7)—, no solo cuando se trata de contenido que incita a la violencia, sino también con otras formas de violencia contra las mujeres. En México, un ejemplo reciente de esto es el caso de la violencia digital que vivieron las estudiantes de la Universidad Anáhuac Mayab, quienes denunciaron un grupo de Telegram en el que se compartían imágenes, videos y datos personales de las estudiantes. Entre las capturas de pantalla que presentaron como evidencia, se encontraba una en la que un usuario señalaba:

> *"creemos otro grupo y ya*
> *(…)*
> *Pero es muy complicado que cierren los grupos, la solución más simple es crear varios respaldos y ya*
> *Inevitablemente cerraran [sic] uno, pero tomará un tiempo hasta que se elimine.*
> *Para ese entonces ya tenemos otros grupos*
> *Y así sin fin*. [12]

9.Efectos de la Violencia Digital

Los efectos negativos del discurso de odio y otros tipos de expresiones dañinas en las mujeres fue documentado por la ONU en los informes ya citados (ONU, 2018 y 2020), adicionalmente, el número de estudios que documentan la evidencia de los efectos negativos de la expresión en

redes sociales crece. Hablando de población general, la prevalencia de discurso de odio y sus efectos es notable. Grupos de población autoidentificados como blancos, afroamericanos y otros reportan tasas altas de incidentes donde sufrieron discriminación en internet, 20%, 29% y 42% respectivamente (Tynes y otros 2008). Tal vez más problemático es que más del 70% de personas autoidentificadas como afroamericanos reportaron haber sido testigos de actos de discriminación en línea contra sus pares. Tynes y otros (2008) pudieron vincular estadísticamente la exposición a discriminación en línea con efectos negativos en la psique de las personas expuestas —mayores niveles de estrés, mayores niveles de infelicidad y disminución de participación en actividades políticas—.

Más problemático aún es que los efectos negativos del discurso de odio en redes sociales han sido vinculados con violencia en la vida real. El caso más conocido es, probablemente, el del papel que se atribuye a Facebook como diseminador de incitación a la violencia entre grupos musulmanes de Sri Lanka. La red social fue señalada por permitir la comunicación entre grupos organizadores y de diseminar los llamados a la violencia hacia las personas que integraban el grupo opositor; la situación escaló tanto que el gobierno del país amenazó con bloquear el acceso a la plataforma (Goel y otros, 2018). El paralelismo entre el caso de Sri Lanka y lo ocurrido en Ruanda durante el genocidio en 1994 es inquietante. De la misma manera que Facebook, las estaciones de radio de Ruanda fueron señaladas como propagadoras de discurso de odio e incitadoras de la violencia. El vínculo real entre el discurso de odio y la violencia genocida en el caso de Ruanda ha sido documentado con detalle (Schabas 2000). Los ejecutivos de las emisoras de radio fueron condenados por un Tribunal Internacional por genocidio (Lafranière 2003).

Los vínculos entre violencia en línea y la vida real para las mujeres comienzan a documentarse con creciente evidencia del impacto negativo en sus vidas. Es difícil relacionar actos de violencia de la vida con posibles incitaciones o provocaciones hechas en línea. Sin embargo, existe evidencia indirecta del peligro en que se pone a las mujeres en el mundo real. El d*oxing*, descrito como una de las violencias comunes en línea, puede causar pánico, ansiedad y alarma. Las víctimas, con frecuencia, deben abandonar sus domicilios o recurrir a medidas preventivas. El 29% de las mujeres en Twitter que experimentan abuso o acoso sufren de *doxing* (Amnistía Internacional, 2018). En un estudio acerca de jóvenes involucrados en pandillas los investigadores encontraron información que vincula ataques a mujeres jóvenes por grupos rivales después de haber cargado contenido a redes sociales (Irwin-Rogers y Pinkney, 2017, p. 8).

Otros estudios han determinado coincidencias estadísticas entre el acceso a internet de alta velocidad en zonas de conflicto social y el incremento de la violencia racial o los crímenes de odio en dichas zonas (Chan y otros, 2016). Así también se ha vinculado el incremento de la violencia contra refugiados y otros crímenes de odio en Estados Unidos durante la presidencia de Donald Trump. El vínculo consider los mensajes del entonces presidente a través de Twitter en donde transmitía sentimientos antirrefugiados y antimusulmanes (Müller y Schwarz 2018; Müller y Schwarz 2021).

Cuando se trata de expresiones en contra de las mujeres, hay algunos estudios que reflejan que las expresiones en redes sociales trascienden el espacio virtual y llevan a las mujeres a actuar para proteger su seguridad. Un estudio de Amnistía Internacional (2018) presenta testimonios de mujeres que han vivido amenazas en Twitter, entre estos, el de una mujer que admitió que había cambiado el apellido a sus hijos en la escuela para que nadie pudiera deducir su relación con ella.

10.Moderación y autogobierno: las redes sociales como entidades de gobierno privadas

Como se ha dicho líneas arriba, las formas en que puede intervenirse en la expresión en redes sociales con el propósito de eliminar o de reducir las expresiones problemáticas y sus efectos son la moderación y la regulación.

Para expandir la definición de Grimmelman (2015) citada antes, ofrecemos la siguiente: por moderación debemos entender la actividad interna que realiza la plataforma de red social, ya sea por motivación propia o a causa de la regulación. Esto significa que la moderación es la serie de criterios y actividades que las plataformas implementan para controlar el contenido que circula por su aplicación. Esto incluye editar, cambiar, suprimir, bloquear, restaurar o prohibir contenido o usuarios.

Todas las plataformas de redes sociales hacen moderación de contenido. Las políticas que informan sus decisiones parten de los criterios generales que tienen de acuerdo con sus objetivos, y de los valores y visión que tienen de su servicio; y se expresan en los términos de uso que las personas usuarias aceptan. Una reflexión interesante al respecto es constatar cuál es el lugar que las plataformas asignan a los valores de la libertad de expresión en sus políticas de moderación. Con frecuencia las políticas de las plataformas son más restrictivas que los criterios internacionales de libertad de expresión. Lo que esto significa es que, con frecuencia, las plataformas prohíben contenido que es perfectamente legal ante los criterios internacionales de libertad de expresión. Probablemente, el mejor ejemplo es el desnudo. Este tipo de expresión está protegido por los valores de la libertad de expresión y es un elemento constante del arte y del entretenimiento para adultos. Pero para lograr que su espacio sea considerado apto para todo público las plataformas prohíben el contenido que incluya desnudos. Estas decisiones varían de plataforma en plataforma, por lo que puede existir una red social en donde el desnudo no esté prohibido, —y las hay, un ejemplo es OnlyFans.com—. Para una plataforma como Facebook, donde privilegian un ambiente "familiar" o al menos no exclusivo para adultos, prohibir los desnudos es una medida adecuada. El problema técnico surge cuando la plataforma debe diseñar los medios para remover el contenido no deseado —desnudo— sin afectar contenido de otro tipo.

Es necesario aclarar que la limitación interna de contenido por parte de una red social —no desnudos en Facebook— no implica vulneración a la libertad de expresión. Esto es porque, por una parte, los usuarios de Facebook han aceptado atenerse a los términos de uso, y aceptaron saber qué tipo de contenidos no son permitidos en la plataforma; así como aceptaron las consecuencias por cargar contenido prohibido, principalmente, la remoción del contenido, pero incluso la cancelación de la cuenta. Por otra parte, Facebook no es la única plataforma de red social, y menos aún el único medio de acceso a la información o a la comunicación, por lo que las personas tienen muchas otras opciones para publicar y encontrar el contenido que les interesa. Este punto, la existencia de una variedad de ofertas y opciones de plataformas, es importante para promover un ambiente de regulación sano en las redes sociales. En general, la consolidación de pocas redes muy grandes limita indirectamente la expresión y acrecienta la importancia de las decisiones de moderación de las redes gigantes (Balkin, 2018; Balkin, 2019). Limitar el acceso a las redes sociales puede lastimar el derecho de acceso a la información y limita el acceso al debate público. En una decisión de la Corte Suprema de los Estados Unidos se expresó que el acceso a las redes sociales es de tal importancia para las personas que las restricciones para su uso deben superar la alta vara del escrutinio estricto (Herrán, 2018).

El problema de la moderación de contenido se vuelve de importancia social cuando vemos el papel que las redes sociales desempeñan en la sociedad actual. Más allá de canales de comunicación o herramientas de contacto, las plataformas se han convertido en parte esencial de la vida de las personas. Autoridades de gobierno de todo el mundo usan las redes sociales para comunicarse efectivamente con millones de personas. Las redes sociales son indispensables en los momentos críticos como los desastres naturales o en los conflictos armados. Permiten la coordinación de autoridades y voluntarios y diseminan información útil. Pero más allá de las grandes aplicaciones sociales, son los pequeños usos y ventajas multiplicados por miles de millones de usuarios que vuelven a las redes sociales en parte importante de nuestras vidas. Información sobre empleos, clima, deportes, entretenimiento, y muchos ejemplos más circulan por las plataformas de redes sociales. El impacto que las redes tienen en las vidas de las personas no puede ser menospreciado. Se constituyen como un elemento esencial de la expresión personal (Klonick 2017) y como parte fundamental de la sociedad democrática moderna (Balkin 2019)[13].

Las redes sociales han mostrado importantes beneficios para la sociedad. Desde la perspectiva política se pueden citar los movimientos que en colectivo son conocidos como la primavera árabe (Tudoroiu, 2014; Howard y otros, 2015). Otros efectos positivos de las redes sociales en la vida y en la sociedad que se han documentado incluyen mejores oportunidades para el desarrollo de redes y trabajos profesionales, acceso a mayor información sobre salud, acceso a mayores oportunidades de empleo, mayor promoción en industrias como el turismo y beneficios para algunas personas con problemas de salud mental (Akram y Kumar, 2017; Buted y otros, 2017; Naslund y otros, 2020). Además de la violencia y del discurso de odio encontramos otros tipos de expresiones problemáticas. La desinformación y las *fake news* son algunos de los problemas recientes que han acaparado atención por sus efectos dañinos. Durante la pandemia ha circulado gran cantidad de información equivocada y engañosa que en el peor de los casos puede poner en peligro la vida de las personas (Cinelli y otros, 2020). En política es común identificar problemas en el abuso de la desinformación por parte de autoridades o de grupos de ciudadanos con fines políticos (Guess y Lyons 2020). Estos son solo algunos ejemplos de los tipos de contenido que circulan por las plataformas y que demandan constante acción de ellas en la forma de moderación de contenido.

11.La Moderación de contenido como autorregulación

Es apropiado reflexionar sobre la necesidad de la moderación de contenido. ¿Pueden las redes sociales no moderar? La respuesta es al mismo tiempo simple y compleja. La versión simple nos recuerda que, aunque la plataforma quisiera mantener una postura completamente neutral ante el contenido generado por terceros, por mandato de ley tiene obligaciones que debe cumplir y responsabilidades en las que puede incurrir, por lo que tiene la necesidad de moderar el contenido en su plataforma. Como ejemplos tenemos que las obligaciones positivas en la plataforma pueden originarse en las prohibiciones del discurso de odio o la apología del terrorismo —caso en Europa—, y las responsabilidades por el contenido que es propiedad intelectual de terceros o por información privada que obtienen de sus usuarios. Esta simple descripción de dos ejemplos muestra que las responsabilidades que tienen las plataformas de comunicación son reales y requieren que estas moderen el contenido. La respuesta complicada requiere agregar las presiones económicas y políticas del mercado, las personas usuarias, los gobiernos y las entidades internacionales.

En general, es posible decir que las principales plataformas de redes sociales —Facebook, Twitter y Youtube— han mantenido una postura flexible y activa en relación con la moderación de contenido. Han modificado sus políticas internas de evaluación, ya sea por presión de las personas usuarias o por amenazas de regulación. Sus decisiones se relacionan con el discurso de odio y contenido discriminatorio, anuncios y propaganda política, desinformación y más. Un ejemplo destacado es el de Facebook —cuya empresa dueña ahora se llama Meta— que creó un órgano externo encargado de evaluar las decisiones de moderación de contenido de la plataforma. Llamado el Consejo Asesor de Contenido, el grupo está formado por reconocidos expertos en libertad de expresión —que incluye juristas, periodistas y activistas, entre otros— y tiene como misión evaluar las decisiones de Facebook bajo los criterios de la libertad de expresión (Klonick, 2020). El Consejo evalúa tanto las decisiones que remueven contenido como las que lo confirman. Sus resoluciones son obligatorias para Facebook y ha tenido importantes decisiones en temas de todo tipo. De manera apropiada para este trabajo resalta la decisión publicada el 28 de enero de 2021 mediante la cual el Consejo revirtió la decisión de Facebook de eliminar en Instagram una publicación que contenía información de salud para concienciar sobre el cáncer de mama durante el "octubre rosa". Un sistema automatizado de Facebook, encargado de hacer cumplir los lineamientos sobre desnudos y actividad sexual de adultos, eliminó la publicación porque contenía imágenes que mostraban "pezones femeninos al descubierto", además de "senos de mujer con los pezones fuera de encuadre o cubiertos con la mano". Facebook reconoció que la decisión fue un error y restauró la publicación.[14] La remoción original —automatizada— por parte de Facebook de las imágenes solo por contener pezones y otras partes del cuerpo de mujeres muestra el sesgo sexista de la moderación y el problema de la sobrerremoción automatizada.

La forma en que las redes sociales moderan contenido no es consistente y las políticas detrás de sus decisiones son cuestionables por los efectos que tienen en grupos vulnerables. Las decisiones de moderación requieren juicios de ponderación entre derechos que entran en conflicto; por una parte, la libertad de expresión y el derecho a la información y por la otra, la privacidad, el honor, la dignidad y más. Balkin (2017; 2018; 2019) ha descrito con claridad las complicaciones que la tecnología de comunicaciones —en específico las redes sociales— han generado en los ejercicios de los derechos individuales. Un análisis de la ponderación y de los derechos implicados en los ejercicios de moderación excede el espacio para este trabajo. Además de la obra de Klonick ya citada (2017) es recomendable la de Keller (2018; 2019; 2020).

A lo largo de los años se ha construido evidencia que indica que las mujeres sufren problemas de imagen personal por publicaciones de redes sociales (Hogue y Mills, 2018; Mills y otros, 2018). Parte de este problema fue confirmado a través de las revelaciones hechas por Frances Haugen, quien fuera ejecutiva en Facebook. En el año 2021 Haugen hizo públicos documentos internos de Facebook en los que se detallaba información producida por la empresa sobre sus productos. Los documentos incluyen investigaciones internas que Facebook no había dado a conocer y que tenían resultados problemáticos para las personas usuarias, especialmente, para las niñas. Durante tres años Facebook condujo investigaciones sobre el impacto de sus productos encontrando que Instagram es dañina para un alto porcentaje de personas usuarias jóvenes, principalmente, mujeres adolescentes (Wells y otros, 2021). Los efectos detectados incluyen problemas como la ansiedad y desórdenes alimenticios, e, inclusive, el pensamiento suicida. Las revelaciones de Haugen dejaron al descubierto las preocupaciones internas de Facebook respecto del impacto de Instagram y otros servicios y del desdén con el que decidieron actuar —o mejor dicho, no actuar— en consecuencia. Es notable que la empresa haya financiado las investigaciones. Lo lamentable es que, aparentemente, haya decidido no hacer públicos los resultados. La duplicidad en la conducta de Facebook es reprochable.

Facebook no es la única plataforma que ha modificado sus políticas de moderación. Estudiar los múltiples cambios realizados por solo las más importantes excedería el espacio disponible para este trabajo. Solo para señalar un ejemplo más de relevancia mencionamos uno de los casos en los que Twitter decidió modificar sus políticas de moderación de contenido. De las principales plataformas de redes sociales, Twitter es, probablemente, la más liberal en lo que libertad de expresión se refiere. Históricamente hacía los menores esfuerzos por incrementar las listas de criterios de expresiones limitadas y buscaba fomentar un espacio libre de debate público. Pero aún bajo estos ideales tuvo que aplicar controles novedosos ante la desinformación propagada en la plataforma, especialmente la que proviene de líderes políticos y jefes de Estado (Alizadeh y otros, 2021), llegando incluso a la prohibición total de la cuenta del presidente Donald Trump.

Como se ha mostrado, el problema de la moderación de contenido es doble; por un lado, tiene un componente político o ideológico. Se trata de responder a las preguntas ¿qué tipo de contenido debe ser moderado? ¿cómo podemos definir con precisión tal contenido? Y, por otra parte, es un problema práctico y técnico, deben diseñarse sistemas de moderación de contenido que sean eficaces, es decir, que eliminen solo el contenido marcado como indeseable. Estos sistemas deben ser mezclas de herramientas automatizadas y revisiones por seres humanos que, bajo un criterio general de transparencia y con respeto a derechos como la revisión y el debido proceso, produzcan decisiones correctas, moderando con éxito las expresiones de la plataforma sin dañar la libertad de expresión y protegiendo la integridad y la dignidad de las personas usuarias. Como podemos imaginar la construcción de un sistema de moderación perfecto es imposible. La tarea urgente es mantener un diálogo constante entre las plataformas de tecnología, las personas usuarias y las y los expertos en comunicación, tecnología y Derecho, de tal manera que puedan diseñarse los sistemas de incentivos más apropiados para el objetivo deseado. Esto es importante, porque se ha documentado que en el mercado actual los incentivos —económicos, sociales y de otros tipos— tienden a resultar en la sobrerremoción de contenido, afectando la libertad y a la sociedad (Keller, 2015).

12.Regulación de internet: ¿un problema insoluble?

Además de la moderación, la otra forma en que se puede controlar la expresión en redes sociales es a través de la regulación. En general, la regulación del internet ha sido descrita como un problema que no tiene solución. La complejidad de la estructura de la red, la variedad de personas con interés —*stakeholders*—, la multiplicidad de factores técnicos y la perspectiva internacional que involucra cientos de Estados con diferentes regímenes jurídicos resulta en que la gobernanza de internet se presenta como un ideal de imposible alcance (Karbulija, 2016).

Distintos Estados han experimentado con maneras diferentes para regular la red. No es posible, o necesario, profundizar en los detalles. Basta mencionar que existen diferencias en la filosofía de regulación: por una parte, encontramos a los Estados que dominan la tecnología de comunicaciones en su interior y que tienen un fuerte control sobre la expresión y sobre el internet. Ejemplos de esto son Corea del Norte, China y, en alguna medida, Turquía. Como contraparte encontramos Estados donde las comunicaciones y la tecnología están abiertas a la libre empresa y a la competencia, donde la iniciativa privada y la competencia participan en el diseño de las políticas de comunicación en gobiernos democráticos. Esto no significa que la libre competencia evite o desconozca la regulación de internet, sino que el proceso democrático y las presiones del mercado permiten la participación de los múltiples interesados (*stakeholders*) en la regulación. En todos los casos —en las versiones restrictivas de los gobiernos autoritarios y en las de libre mercado—, la regulación se aplica de manera directa, por ejemplo, a través de

legislación; e indirecta, como en el caso de políticas públicas o presiones políticas, como el caso del *jawboning* (Keller, 2019).

La regulación directa es la que afecta a los componentes de la red de manera inmediata. Esta puede ser técnica y aplicable a elementos de infraestructura, por ejemplo, la determinación que impide el acceso a un satélite, una parte de espectro de transmisión o algo similar. O pueden ser normas que establezcan políticas específicas de comunicación u obligaciones determinadas para los involucrados. Como ejemplo podemos citar las políticas de neutralidad de la red que una nación puede adoptar o las reglas que indican los requisitos que deben cumplir los proveedores de servicios de internet, entre otros.

Puede verse entonces que el campo de acción de la regulación es muy amplio y, con frecuencia, las acciones realizadas dentro de un área tienen efectos no esperados en otra. Esto es parte de lo que vuelve complicado plantear estrategias de regulación a gran escala. Pero en el caso de las redes sociales la regulación normalmente es discutida en la forma de normas jurídicas que obliguen a las plataformas de redes sociales a comportarse de alguna manera determinada. La historia de la regulación de las redes sociales, y de internet, porque se desarrollaron juntas, es complicada, involucra las decisiones del sistema jurídico de Estados Unidos en los años 90 del siglo pasado y los criterios cambiantes de otros países —principalmente, en Europa— en este siglo.

13.Regulación de internet en Estados Unidos y Europa

En Estados Unidos, en 1996, se creó legislación que protege a los proveedores de servicios interactivos por computadoras —esto incluye no solo a las redes sociales, sino a todos los servicios digitales por internet— inmunizándolos de la responsabilidad por el contenido proporcionado por terceros. Los autores de la ley, conocida comúnmente como la sección 230 de la CDA, la crearon con el propósito expreso de que los proveedores —en nuestro caso redes sociales— pudieran moderar[15] contenido libremente sin el temor de que fueran responsabilizados por el mismo. La ley deja en claro que las personas que entregan el contenido a los proveedores son los únicos responsables de este. De esta manera en Estados Unidos se creó un poderoso escudo para las empresas de internet. La jurisprudencia solidificó la protección de la sección 230 en ese país y actualmente es casi imposible lograr una responsabilidad por parte de una red social o de otra empresa proveedora de información por internet por el contenido creado por terceros. Este no es un tema menor. Sin esta protección páginas como Youtube, Yelp y Amazon no podrían existir.

Jeff Kossef, en su libro *Las 26 palabras que crearon el interne*t (2019, p. 141-143) en el que da cuenta de la historia de la legislación y la jurisprudencia relacionada con la sección 230 narra la decisión de un importante caso sobre abuso sexual de una niña relacionado con una red social.[16] Julie Doe —un seudónimo— tenía 13 años cuando abrió una cuenta en Myspace en 2005. Un año después un hombre de 19 años, Pete Solis, se puso en contacto con ella a través de la red social y comenzaron a enviarse mensajes. Ella le dio su número de teléfono, finalmente se conocieron en persona y Solis la atacó sexualmente. La niña informó a su madre y ésta, a la policía. Solis fue arrestado por delitos sexuales. La madre de la niña demandó a Myspace por negligencia y fraude, entre otros, por su papel negligente al no desarrollar medidas preventivas que impidieran el trágico ataque a la niña. El litigio se desarrolló en la corte federal de Texas y la decisión final fue tomada por la Corte de Apelaciones del Quinto Circuito. La Corte de Circuito decidió a favor de Myspace con base en la protección que la sección 230 le otorgaba, Kossef cita la sentencia: *Sus pretensiones están impedidas por la CDA*. Sobre el impacto del caso hacia el futuro, Kossef escribe:

> *[La] decisión generó una sentencia de apelación federal obligatoria en la que clarificaba que la sección 230 aplicaba a las compañías de redes sociales. Otras cortes de la nación la citarían después, en sus sentencias inmunizando a otras compañías de redes sociales como Facebook y Twitte*r (p. 142).

En los años posteriores a la legislación americana y a su jurisprudencia la mayoría de los países siguieron el mismo camino, probablemente por la influencia de los Estados Unidos en el mundo tecnológico, pero también porque la idea de inmunizar al proveedor de servicios por el contenido de terceros es buena. En Europa, en el año 2000 se adoptó la Directiva de Comercio Electrónico, que incluye un régimen de responsabilidad limitada e inmunidad para los servicios que solo funcionan como conductos. Sin embargo, con el paso de los años el continente europeo se ha desviado de esta idea de no intervención e inmunidad y ha creado, poco a poco, importantes regulaciones para las plataformas de redes sociales.

El caso más importante es el de la legislación de Alemania conocida como Ley de Control de la Red, o NetzDG. Esta legislación contiene importantes obligaciones para las plataformas como la de remover contenido rápidamente una vez que ha sido notificada por la usuaria, e incluye fuertes sanciones y multas por incumplimiento. Una de sus virtudes es la obligación de transparencia, mediante la cual se ordena a las plataformas publicar periódicamente informes detallados sobre los casos y las decisiones relacionadas con la moderación de contenido. Aunque en general la ley ha sido criticada por ser inadecuada y por crear condiciones poco apropiadas para el ejercicio de la libertad de expresión, sus obligaciones sobre transparencia se han convertido en un ejemplo de obligaciones útiles e importantes para promover la moderación de contenido.

La NetzDG ha sido copiada por países como Rusia y Turquía que no tienen los mismos compromisos con la libertad expresión que Alemania, elevando la preocupación de que la regulación pueda ser usada por Estados totalitarios para acallar la expresión. Además de la ley alemana existen otras importantes regulaciones que afectan a las plataformas de redes sociales en Europa. Destacan el Reglamento General de Protección de Datos de la Unión Europea, que regula la información privada y la manera en que las empresas deben relacionarse con ella, y el derecho al olvido, consistente en la obligación de motores de búsqueda de desindexar de sus resultados aquellos que ya no son relevantes para la sociedad y que pueden causar daño. Para acabar, actualmente se discute en el Parlamento Europeo la Ley de Servicios Digitales, que se propone regular específicamente las plataformas de redes sociales. Representa una importante oportunidad para crear un sistema normativo eficiente y efectivo que considere los beneficios de las plataformas y proponga herramientas útiles para su control.

Conclusiones

Lo expuesto en este trabajo muestra que el panorama de regulación de redes sociales en general es complicado y cuando se involucran los derechos de las personas en situación de vulnerabilidad y los derechos de las mujeres, puede resultar aún más complicado. Sin embargo, es necesaria una mirada crítica a las redes sociales desde la perspectiva de género.

Indudablemente, la regulación del contenido en las redes sociales es indispensable y se han hecho diversos esfuerzos, tanto de manera interna como externa, por lograrlo. Sin embargo,

para que la regulación beneficie y proteja a todas las personas usuarias, se requiere incluir de manera permanente una visión de derechos humanos y de género que pueda considerar las necesidades particulares de cada grupo, —especialmente los grupos en situación de vulnerabilidad— para diseñar acciones afirmativas que permitan el acceso y disfrute de estas poderosas herramientas de comunicación en condiciones de igualdad y libres de violencia.

Es patente que aún queda mucho por hacer para erradicar la violencia contra las mujeres en las redes sociales y que se requiere que muchos actores se involucren en la regulación de estas plataformas. Es un llamado para que los dueños y responsables de las plataformas de redes sociales, proveedores de servicios de internet, organizaciones de la sociedad civil, organismos internacionales, autoridades locales, las propias mujeres y también los hombres, nos comprometamos a pasar a la acción para que las redes sociales sean efectivamente espacios libres y seguros para todas las personas usuarias.

---

[1] Organización de Naciones Unidas, Resolución 71/199: El derecho a la privacidad en la era digital, aprobada por la Asamblea General el 19 de diciembre de 2016.

[2] Consiste en investigar y divulgar información de carácter personal sobre una persona sin su consentimiento. En el caso específico de las mujeres, muchas veces se publica información personal de la víctima insinuando que está ofreciendo servicios sexuales

[3] Así se conoce a la técnica con la cual se busca conseguir información de las personas a través del engaño, ganándose la confianza de la víctima al hacerse pasar por una persona, empresa o servicio de confianza para manipularla y hacer que realice acciones que no debería realizar, por ejemplo, revelar contraseñas o hacer *click* en un enlace.

[4] Vale la pena aclarar que el *sexting* no es una forma de violencia, sino una práctica común entre parejas. Es parte de la libertad sexual de las personas y está protegida por el derecho a la intimidad y a la vida privada.

[5] Al conjunto de reformas en esta materia se le conoce como “Ley Olimpia”.

[6] Usaremos la expresión “discurso de odio” como término general que abarque a todos los tipos de expresiones que por diversas causas pueden ser consideradas discriminatorias.

[7] Esto no significa que no existan restricciones constitucionales basadas en el contenido. La Corte Suprema de Estados Unidos ha delimitado varias áreas de expresión restringida. Algunas de estas son categorías de restricción absoluta como la pornografía infantil, las palabras violentas (*fighting words*) y las amenazas verdaderas.

[8] Brandenburg v. Ohio, 395 U.S. 444 (1969)

[9] Esta actividad consiste en identificar el contenido y compararlo con una base de datos de contenidos sospechosos o prohibidos. Las políticas de moderación determinarán qué contenido debe incluirse en la base. Al identificar el contenido publicado por la persona usuaria y determinar que tiene un par en la base de datos, el sistema procederá a realizar la acción de moderación indicada (bloqueo, imposición de etiqueta, etc.). El sistema de identificación puede analizar palabras, textos, símbolos e incluso imágenes o videos usando aplicaciones de *hasheo*.

[10] Esta es una descripción simplificada de las actividades básicas de moderación, pero es suficiente para los propósitos de este artículo. Para adentrarse en las particularidades técnicas del funcionamiento se recomienda ver a Grimmelman, 2020, ya citado.

[11] Para más información sobre las categorías sospechosas ver BAYEFSKY, A. F. *El principio de igualdad o no discriminación en el derecho internacional*, 2016.

[12] Para más información sobre este caso se pueden consultar los siguientes enlaces: https://www.hazruido.mx/reportes/violencia-digital-en-la-anahuac-mayab-estudiantes-difunden-fotos-y-videos-intimos-de-sus-companeras/ (consultado el 14 de febrero de 2022). https://vocesfeministas.mx/no-hay-avances-en-investigaciones-sobre-el-caso-anahuac-de-violencia-digital-contra-mujeres/ (consultado el 14 de febrero de 2022).

[13] No pasa desapercibido el problema de la desinformación y las *fake news* en las redes sociales. Es un problema importante que se relaciona también con la moderación y la regulación. Sin embargo, no se profundiza en ello porque distraería del punto central que es el estudio de la problemática de la violencia contra grupos vulnerables en redes sociales.

[14] La decisión puede ser consultada en https://oversightboard.com/decision/IG-7THR3SI1/ (consultado el 14 de febrero de 2022).

[15] Es un error común pensar que la sección 230 de la CDA otorga inmunidad solo para alojar contenido, pero el texto de la sección (2) de la ley es claro cuando otorga inmunidad por las actividades de moderación de buena fe. En un blog publicado en el 2020 Chris Cox —ex representante de California en el Congreso de los Estados Unidos (R 1998-2005) y uno de los autores de la ley— escribió:

> *Un estándar legal que protegiera solo los sitios web donde "todo vale" de la responsabilidad ilimitada por el contenido generado por el usuario habría sido un duro golpe para el Internet. El representante Wyden y yo decidimos que la moderación de contenido de buena fe no debería ser castigada* (Cox, 2020).

El texto íntegro de las secciones (1) y (2) de la ley dice en su original (47 U.S. Code § 230):

> *(1)Treatment of publisher or speaker.No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider. (2)Civil liability No provider or user of an interactive computer service shall be held liable on account of—(A)* ***any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected****; or (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).* (énfasis añadido, subrayado de los autores).

[16] Doe v. MySpace Inc. 28 F.3d 413 (5th Cir. 2008).

Referencias

AKRAM, W. & KUMAR, R. (2017). "A study on positive and negative effects of social media on society". *International Journal of Computer Sciences and Engineering,* 2017, 5(10), 351-354.

ALIZADEH, M., GILARDI, F., HOES, E., KLÜSER, K.J. y otros, (2021). *Content moderation as a political issue: The twitter discourse around Trump's ban.* University of Zurich.

AMNISTÍA INTERNACIONAL. (2022). Amnesty reveals alarming impact of online abuse against women. In., 2017.

AMNISTÍA INTERNACIONAL. #TOXICTWITTER. 2018. Violencia y abuso contra las mujeres en internet.

BALKIN, J. M. (2017). "Free speech in the algorithmic society: big data, private governance, and new school speech regulation". *UCDL Rev*., 51, 1149.

BALKIN, J. M. (2018)."Free Speech is a Triangle". *Columbia Law Review*, 118(7), 2011-2056.

BALKIN, J. M. (2019). *How to Regulate (and Not Regulate) Social Media.*

BEN-DAVID, A. & MATAMOROS-FERNÁNDEZ, A. (2016). "Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain". *International Journal of Communication,* 10, 27.

BLEICH, E. & AL-MATEEN, S. (2021). "Hate Speech and the European Court of Human Rights: Ideas and Judicial Decision-Making". *Mich. St. Int'l L. Rev.,* 29, 179.

BUTED, D. R. y otros. (2014). "Effects of social media in the tourism industry of Batangas Province". *Asia Pacific Journal of Multidisciplinary Research* vol. 2, no 3.

CAMPOS, M., RAMOS, M., TREJO DELARBRE, R., HERNÁNDEZ RAMÍREZ, M.E. y otros. (2015). *Mensajes de odio y discriminación en las redes sociales*. México: CONAPRED. ISBN 978-607-8418-10-7.

CARR, C. T. & HAYES, R.A. (2015). "Social media: Defining, developing, and divining". *Atlantic journal of communication,* 23(1), 46-65.

CHAN, J., GHOSE, A. & SEAMANS, R. (2016). The internet and racial hate crime: Offline spillovers from online access. *Mis Quarterly,* 40(2), 381-403.

CHANDRASEKHARAN, E., PAVALANATHAN, U. , SRINIVASAN, A., GLYNN, A. y otros. (2017). "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech". *Proceedings of the ACM on Human-Computer Interaction.* 1(CSCW), 1-22.

CDHCM. Comisión de Derechos Humanos de la Ciudad de México. Violencia Digital Contra las Mujeres en la Ciudad de México. 2021.

CINELLI, M., QUATTROCIOCCHI, W., GALEAZZI, A., VALENSISE, C. M. y otros. (2020). The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004.

COX, C. (2020). "The Origins and Original Intent of Section 230 of the Communications Decency Act". *Journal of Law and Technology Blog, University of Virginia*. Disponible en: https://jolt.richmond.edu/2020/08/27/the-origins-and-original-intent-of-section-230-of-the-communications-decency-act/ (consultado el 14 de febrero de 2022).

DEEPTRACE. *The State of Deepfakes: Landscape, Threats, and Impact.* 2019.

DUARTE, N., LLANSO, E. & LOUP, A.C. (2018). "Mixed Messages? The Limits of Automated Social Media Content Analysis". In *2018 Conference on Fairness, Accountability, and Transparency* , p. 106.

ELSHERIEF, M., NILIZADEH, S., NGUYEN, D., VIGNA, G. y otros. (2018). "Peer to peer hate: Hate speech instigators and their targets". In *Proceedings of the International AAAI Conference on Web and Social Media.* vol. 12.

ESQUIVEL ALONSO, Y. (2016). "El discurso del odio en la jurisprudencia del Tribunal Europeo de Derechos Humanos". *Cuestiones constitucionales,* (35), 3-44.

FLORES, C., FLORES, C., GUASCO, T., & LEÓN-ACURIO, J. (2017). "A diagnosis of threat vulnerability and risk as it relates to the use of social media sites when utilized by adolescent

students enrolled at the Urban Center of Canton Cañar". *International Conference on Technology Trends.*

FRANKS, M. A. *Drafting an effective'revenge porn'law: A guide for legislators*. Available at SSRN 2468823, 2015.

GARCÍA ROMÁN, M. & MINDEK JAGIC, D. (2021). "Ciberviolencia de género en redes sociales". *Controversias y Concurrencias Latinoamericanas*, 2021, 12(22), 333-349.

GOEL, V., KUMAR, H. & FRENKEL, S. (2018)." In Sri Lanka, Facebook Contends With Shutdown After Mob Violence". In *The New York Times.*

GORWA, R., BINNS, R. & KATZENBACH, C. (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". *Big Data & Society*, 7(1), 2053951719897945.

GRIMMELMAN, J. (2015)."The Virtues of Moderation". *Yale Journal of Law and Technology*, 42.

GUESS, A. M. & LYONS, B. A.. "Misinformation, disinformation, and online propaganda". In N. PERSILY AND J.A. TUCKER (eds.) *Social Media and Democracy: The State of the Field, Prospects for Reform.* Cambridge UK: Cambridge University Press, p. 10-33.

HARRIS, B. & L. VITIS (2020). "Digital intrusions: technology, spatiality and violence against women". *Journal of gender-based violence*, 4(3), 325-341.

HEREDIA, A. V. (2017) "Los discursos del odio. un estudio jurisprudenciaL/The hate speeches. A jurisprudential study". *Revista española de Derecho Constitucional,* (110), 305-334.

HERNÁNDEZ, M. D. (2020). *Discurso de odio en América Latina*. Derechos Digitales.

HOGUE, J. V., MILLS, J. S. (2019)."The effects of active social media engagement with peers on body image in young women". *Body image*, vol. 28, p. 1-5.

HERRÁN AGUIRRE, A.(2018). "Derecho a la libertad de expresión y acceso a las redes sociales: El caso Packingham con Carolina del Norte". *Revista Chilena de Derecho y Tecnología.* 7(2), 163-186.

HOWARD, P. N., DUFFY, A., FREELON,D., HUSSAIN, M. y otros. (2015). "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" *SSRN Electronic Journal*.

KAAKINEN, M., RÄSÄNEN,P. , NÄSI, M., MINKKINEN, J. y otros (2018). "Social capital and online hate production: A four country survey". *Crime, Law and Social Change*, 69(1), 25-39.

KARBULIJA, J. (2016). *Introducción a la gobernanza de internet*. Edtion ed.: Diplo. ISBN 978-99932-53-31-0.

KAVANAGH, E., LITCHFIELD, C. & OSBORNE, J. (2019). "Sporting women and social media: Sexualization, misogyny, and gender-based violence in online spaces". *International Journal of Sport Communication,* 12(4), 552-572.

KELLER, D. (2015). *Empirical Evidence of 'Over-Removal' by Internet Companies under Intermediary Liability Laws*. The Center for Internet and Society at Stanford Law School.

KELLER, D. (2018). *Toward a Clearer Conversation About Platform Liability.* Knight First Amendment Institute's "Emerging Threats" essay serie.

KELLER, D. (2019). "Who Do You Sue?" *Hoover Institution Aegis Series Paper,* (19002).

KELLER, D. (2020). "Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling". *GRUR International,* 69(6), 616-623.

KLONICK, K. (2017). "The new governors: The people, rules, and processes governing online speech". *Harv. L. Rev.,* 131, 1598.

KLONICK, K. (2020). "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression". *The Yale Law Journal.* 129(8), 2232-2605.

KOSSEFF, J. (2019). *The twenty-six words that created the internet*. Edtion ed.: Cornell University Press. ISBN 1501735780.

LAFRANIERE, S.(2003)."Court Finds Rwanda Media Executives Guilty of Genocide". In *The New York Times.*

MALMASI, S. & M. ZAMPIERI, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

MILLS, J. S., y otros (2018). "*Selfie* harm: Effects on mood and body image in young women". *Body image*, vol. 27, p. 86-92.

MÜLLER, K. & SCHWARZ, C. (2018). "Making America hate again? Twitter and hate crime under Trump. Unpublished working paper". University of Warwick.

MÜLLER, K. AND SCHWARZ, C. (2021). "Fanning the flames of hate: Social media and hate crime". *Journal of the European Economic Association,* 19(4), 2131-2167.

NASLUND, J. A., BONDRE, A., TOROUS, J. & ASCHBRENNER, K.A. (2020). "Social media and mental health: benefits, risks, and opportunities for research and practice". J*ournal of technology in behavioral science,* 5(3), 245-257.

OEA. (2021). La violencia de género en línea contra mujeres y niñas. Guía de conceptos básicos, herramientas de seguridad digital y estrategias de respuesta.

ONU. (2018). Informe de la Relatora Especial sobre la violencia contra la mujer, sus causas y consecuencias acerca de la violencia en línea contra las mujeres y las niñas desde la perspectiva de los derechos humanos.

ONU MUJERES. (2020). Violencia contra mujeres y niñas en el espacio digital: Lo que es virtual también es real.

PLAN INTERNATIONAL. (2020). *Free to be online? Girls' and young women's experiencies of online harassment.*

RELIA, K., LI, Z., COOK, S.H. & CHUNARA, R. (2019). "Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities". In *Proceedings of the International AAAI Conference on Web and Social Media.* vol. 13, p. 417-427.

SCHABAS, W. A. (2000). "Hate speech in Rwanda: The road to genocide". *McGill LJ,* 46, 141.

SELLARS, A. (2016). "Defining hate speech". *Berkman Klein Center Research Publication*, (2016-20), 16-48.

SIEGEL, A. A. (2020). Online hate speech. In *Social Media and Democracy: The State of the Field, Prospects for Reform.* Cambridge, UK: Cambridge Univesity Press, p. 56-88.

STEPHAN III, P. B. (1982)."The First Amendment and Content Discrimination". *Virginia Law Review,* 203-251.

TUDOROIU, T. (2014) "Social media and revolutionary waves: The case of the Arab spring". *New Political Science,* 36(3), 346-365.

TWENGE, J. M., HAIDT, J., LOZANO, J. & CUMMINS, K.M. (2022). "Specification curve analysis shows that social media use is linked to poor mental health, especially among girls". *Acta psychologica,* 224, 103512.

TYNES, B. M., GIANG,M.T., WILLIAMS, D.R. & THOMPSON, G.N.(2008)."Online racial discrimination and psychological adjustment among adolescents". *Journal of Adolescent Health,* 43(6), 565-569.

UNESCO. (2015). Broadband commission for digital development working group on broadband and gender, Cyber violence against women and girls: A world-wide wake-up call.

VERGÉS BOSCH, N., HACHE, A., MANZANARES REYES, G., ESCOBAR, M. M. y otros (2017). *Redes sociales en perspectiva de género: guía para conocer y contrarrestar las violencias de género on-line*. Edtion ed.: Instituto Andaluz de Administración Pública, ISBN 8483336839.

VILLANUEVA, D. N., FECED, S.C., CALVO, B.R. & BARRANCO, I.B. (2017). "Influencia negativa de las redes sociales en la salud de adolescentes y adultos jóvenes: una revisión bibliográfica". *Psicología y salud,* 27(2), 255-267.

WELLS, G., HORWITZ, J. & SEETHARAMA, D. (2021)." Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show". In *The Wall Street Journal.*

WOODLOCK, D., MCKENZIE, M., WESTERN, D. & HARRIS, B. (2020). Technology as a weapon in domestic violence: Responding to digital coercive control. *Australian social work,* 73(3), 368-380.