Propuesta para delimitar el concepto de "contenido dañino en línea" a partir de la experiencia del Reino Unido

Proposal to define the concept of "harmful online content" based on the experience of the United Kingdom

Jorge Tisné Niemann*

Universidad de los Andes, Santiago, Chile
jtisne@bofillmir.cl | https://orcid.org/0000-0003-2651-2648

◆

Recibido: 10/06/2024 | Aceptado: 29/07/2024 | Publicado: 07/08/2024

Resumen: El uso masificado de internet ha provocado que los usuarios estén expuestos a una gran cantidad de conductas ofensivas por parte de terceros, lo que suele incidir en la vulneración de los derechos y libertades de las personas. Con el objeto de promover un espacio seguro, distintas iniciativas legislativas comparadas han comenzado a regular los daños en línea, exigiendo a las plataformas ciertos deberes de moderación de dicho contenido. Sin embargo, el concepto de daño en línea es subjetivo, especialmente si se considera que incluye aquel contenido que sin ser ilegal, es considerado nocivo o perjudicial. A propósito de lo anterior, este trabajo tiene por objeto estudiar las distintas conductas que pueden subsumirse en el concepto de daño legal pero perjudicial a la luz de la taxonomía de daños ofrecida en el Libro Blanco de Daños en Línea del Reino Unido, para luego proponer una definición de daño en línea que sirva de base para una regulación orgánica y coherente de las conductas que deben ser atenidas y controlados en entornos digitales.

Palabras clave: Internet; daño en línea; conductas legales pero perjudiciales.

Abstract: The massive use of the Internet has exposed users to a large amount of offensive conducts of third parties, directly affecting the rights and freedoms of individuals. In order to promote a online safe space, different comparative legislative initiatives have begun to tackle online harms, requiring certain duties of moderation from platforms. However,

* Abogado, Magíster en Investigación Jurídica y Doctor en Derecho, todos en la Universidad de los Andes, Chile. Law Master in Innovation, Technology and the Law, Universidad de Edimburgo, Escocia. Director del Diplomado en Derecho y Tecnología de la Universidad de los Andes / Eclass. Asociado senior en el área de IP, datos y tecnología en Bofill Mir Abogados.

the concept of online harm is subjective, especially if we consider that it includes content that is not illegal but is considered harmful or damaging. In view of the above, this paper aims to study the different conducts that can be subsumed under the concept of legal but harmful content in the light of the taxonomy of harms offered in the UK White Paper on Online Harm and then propose a definition of online harm that serves as a basis for an organic and coherent regulation of the conducts that must be mitigated and controlled in digital environments.

Keywords: Internet; online harm; legal but harmful conduct.

1. Introducción

El uso masificado de internet ha traído consigo innumerables beneficios para los usuarios. En todos los ámbitos de la vida se pueden observar avances a propósito del mayor acceso a información instantánea, herramientas que promueven un trabajo más eficiente, e incluso, procesamiento de datos masificados que junto con la inteligencia artificial admiten alcanzar conocimientos y desarrollos que se estima redefinirán el mundo que vivimos.

Sin embargo, junto con estos beneficios, internet presenta un importante número de desafíos que aún no han logrado ser abordados de manera adecuada o uniforme en las distintas jurisdicciones. Uno de esos desafíos, y que a nuestro juicio resulta ser uno de los más complejos de resolver, es definir qué contenido en línea debe ser identificado como indeseado, y por lo tanto, prohibido para su circulación. Esto es relevante, especialmente en sociedades donde el uso y acceso masificado de internet es un bien jurídico considerado digno de protección, incluso a nivel constitucional[1].

En este sentido, el contenido dañino en línea genera graves efectos perjudiciales, ya que se propaga instantáneamente, afecta a una cantidad indeterminada de personas, permanece visible durante años en internet y a menudo son generados de manera anónima. Al igual que los comportamientos realizados en el mundo material, los daños en línea deben abordarse de manera adecuada para proteger a los usuarios y garantizar que internet continúe siendo un espacio seguro.

Con todo, encontrar el equilibrio correcto entre el contenido indeseado y la libertad de expresión e información es una tarea particularmente difícil. Existe una amplia gama de daños en línea que generalmente involucran términos indefinidos y subjetivos, y deben evaluarse en relación con elementos igualmente inciertos, tales como consideraciones personales, tiempo de exposición, contexto, historia e incluso idiosincrasia del sujeto afectado. Asimismo,

[1] A modo de ejemplo, el artículo 86 de la Propuesta de Nueva Constitución, rechaza en el plebiscito del 4 de septiembre de 2023, consagraba el derecho al acceso universal a la conectividad digital y a las tecnologías de la información y comunicación. En ese sentido, ver Prince Torres, 2020.

el esfuerzo de definir un concepto de contenido dañino sobrepasa la descripción de meras conductas ilegales, pues la controversia más desafiante, y de la cual se ocupa este trabajo, tiene que ver precisamente con aquellos contenidos o comportamientos que no infringen disposiciones legales, pero que podrían considerarse de todas maneras perjudiciales e indeseados para la sociedad y para ciertos sujetos en particular.

La delimitación del concepto que se propone estudiar en el presente trabajo es particularmente relevante pues internet es un entorno altamente dinámico y caracterizado por una constante evolución.

En este sentido, la falta de un concepto claro que oriente las nuevas regulaciones que intenten controlar el contenido dañino en línea, genera un escenario en donde medidas aisladas o fragmentadas pueden afectar la innovación tecnológica y la libertad de expresión, obligando a los proveedores de servicios de internet (en adelante "ISP" por sus siglas en inglés) a adoptar potentes sistemas de moderación de contenido, que redundarían posiblemente en mecanismos de censura para los usuarios. Por el contrario, un ambiente absolutamente desregulado, en donde todo contenido fuera permitido, desfavorecería la participación de los usuarios en internet y socavaría la confianza en este espacio de interacción.

Dado que en Chile no existe un esfuerzo sistemático por abordar el asunto, es interesante conocer la iniciativa del gobierno del Reino Unido para controlar los daños en línea, el cual comenzó en el año 2019 con el denominado Libro Blanco de Daños en Línea[2] y que concluyó en octubre de 2023 con la publicación de la Ley de Seguridad en Línea[3]. Si bien finalmente el *Online Safty Act* restringió de manera importante las conductas en línea reguladas, el OHWP originó un intenso debate a propósito de la intención del Reino Unido de regular el contenido ilegal y también el legal pero indeseado (también conocido como contenido nocivo). Parte de esta discusión servirá de base para orientar los lineamientos que presentaremos en este trabajo.

El OHWP, entre otros elementos, fue una propuesta pretensiosa pues tuvo por finalidad regular, desde una perspectiva integral y sistemática, una amplia lista de contenidos perjudiciales e implementar un deber de cuidado para los intermediarios de internet o ISP. Además, cuando se trata de regulaciones en entornos digitales, usualmente se carece de un sistema regulatorio uniforme y coherente que agrupe dichas conductas. Este problema se observa en nuestro país, pues existe un escenario regulatorio fragmentado que evidencia la necesidad de contribuir a delimitar una concepción del daño en línea que sirva para orientar la futura política pública en la materia.

En ese sentido, algunas de las conductas identificadas en el extranjero como dañinas en línea encuentran en nuestro país cierto grado de regulación, dispersa en distintas normativas. Por ejemplo, es posible identificar que sexting, explotación y abuso sexual

[2] En adelante OHWP por sus siglas en inglés.

[3] En adelante se usará el título original *Online Safety Act*.

infantil ha sido recogido en la ley N° 21.522 de 2022 que modificó el Código Penal. A su vez, el contenido y actividad terrorista se encuentra regulado en la ley N° 18.314 de 1984. La incitación a la violencia se encuentra sancionada en el artículo 31 de la ley N° 19.733, sobre Libertad de opinión e Información y Ejercicio del Periodismo. Asimismo, actualmente existe en el Congreso un proyecto de ley (Boletín N° 12.164-07) que tiene por finalidad sancionar la pornografía de venganza. Junto con lo anterior, se debe mencionar la Comisión Asesora contra la Desinformación, creada mediante Decreto Supremo Nº 12, de fecha 12 de mayo de 2023 y cuyo término se decretó mediante Decreto Supremo N° 5 de fecha 6 de marzo de 2024, ambos del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación.

La hipótesis de este artículo es que es posible esbozar los lineamientos generales de una definición de contenido dañino en línea que permita, con posterioridad, concebir una regulación adecuada y proporcional respecto al control del contenido publicado en entornos digitales, especialmente a propósito de la diligencia que se exija de los ISP para su prevención. Para efectos metodológicos, se utilizará la toxonomía de daños prevista en el OHWP como catálogo determinado de conductas que podrían ser subsumidos en el concepto que se intenta definir.

Este trabajo se divide en las siguientes secciones. En primer lugar, se abordará de manera concisa los daños bajo estudio y el deber de cuidado previsto en el OHWP. La segunda sección presentará la primera distinción entre contenidos ilegales y contenido legales pero perjudiciales. La tercera sección se centrará en los contenidos legales pero perjudiciales, en donde se discutirán los problemas relacionados con su definición. En la cuarta sección se abordarán ciertos elementos del discurso de odio que se proponen como punto de partida para delimitar los contornes del contenido legal pero perjudicial en línea. En base a lo anterior, en la quinta sección se utilizarán dichos elementos del discurso de odio para categorizar en un esquema ordenado y coherente las distintas conductas identificadas por el OWHP. En virtud del análisis previo, en la sexta sección se concluirá con la proposición de un concepto de contenido dañino en línea.

2. Breve descripción del *Libro Blanco de Daños en Línea* y el *Online Safety Act* del Reino Unido

En abril de 2019, el Secretario de Estado para la Cultura, Medios de Comunicación y Deporte, en conjunto con el Secretario de Estado de Interior, ambos del Reino Unido, presentaron el OHWP. Mediante este documento, el gobierno declaró su intención de consolidar al Reino Unido como "el lugar más seguro del mundo para navegar en línea y el mejor lugar para comenzar y hacer crecer un negocio digital"[4].

[4] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

A la luz de este propósito, el OHWP establece un marco regulatorio para abordar la actividad en línea ilegal, y asimismo, la actividad en línea indeseada. Durante febrero de 2020, concluyó la etapa de respuesta inicial a la consulta del OHWP, con el gobierno del Reino Unido proporcionando respuestas a las principales preocupaciones de las partes interesadas. Luego, en diciembre de 2020, el gobierno del Reino Unido publicó la respuesta definitiva y complementaría a la consulta sobre el OWHP. En base a esta discusión previa, se presentó al Parlamento del Reino Unido el proyecto de Ley sobre Seguridad en Línea (*Online Safety Act*), la cual fue aprobada recientemente en octubre de 2023 por el Parlamento Británico.

Los ejes principales del OHWP fueron proporcionar claridad para las compañías, un nuevo marco regulatorio para los daños en línea, definir las compañías sujetas a las nuevas obligaciones, facultades de la agencia reguladora, exigibilidad del nuevo régimen regulatorio y el uso de la tecnología para promover soluciones (OHWP, pp. 6-10). Para efectos de este artículo, solo se abordará el tema relacionado con los daños objeto de la regulación.

El gobierno del Reino Unido tuvo como motivación para la discusión del OHWP el hecho de que existen una serie de iniciativas legales y otras de carácter voluntarias por parte de los ISP para abordar los daños en línea, pero no han sido lo suficientemente efectivas para detener su propagación en internet, especialmente aquellos que amenazan a los niños. La iniciativa planteó una amplia gama inicial de daños dentro del alcance de la propuesta, clasificados en tres categorías diferentes: daños con definiciones claras, daños con una definición menos clara y exposición de menores a contenido legal.

La lista de daños con una definición clara incluye: (i) explotación y abuso sexual infantil, (ii) contenidos y actividades terroristas, (iii) delincuencia organizada de inmigración, (iv) esclavitud moderna, (v) pornografía extrema, (vi) pornografía vengativa, (vii) acoso y ciberacoso, (viii) delitos motivados por el odio, (ix) incitación o ayuda al suicidio, (x) incitación a la violencia, (xi) venta de bienes/servicios ilegales, como drogas y armas (en internet abierto), (xii) contenidos subidos ilegalmente desde prisiones, y (xiii) sexting de imágenes indecentes por menores de 18 años.

La lista de daños con una definición menos clara incluye: (i) ciberacoso y trolling, (ii) contenido y actividad extremista, (iii) comportamiento coercitivo (iv) intimidación, (v) desinformación, (vi) contenido violento, (vii) apología de la autolesión y (viii) promoción de la mutilación genital femenina (MGF).

La lista vinculada con la exposición de menores a contenidos legales incluye: (i) niños que acceden a pornografía, y (ii) niños que acceden a material inapropiado (incluidos menores de 13 años que usan redes sociales, menores de 18 años que usan aplicaciones de citas y tiempo de pantalla excesivo).

Como se observa, la regulación contemplaba una variada y disímil cantidad de tipos de daños, lo cual hizo del OHWP una iniciativa particularmente compleja y controvertida en el Reino Unido. Lo anterior, pues no solo tenía por objeto regular daños en línea que

pudieran ser clasificados como ilegales, sino que intentó abordar conductas que no tenían una definición específica, y en consecuencia, quedaría a la discreción de los ISP la prevención de dichos contenidos legales pero dañinos.

El OHWP reconoció que un grupo relativamente pequeño de empresas ha participado en esfuerzos voluntarios liderados por el gobierno del Reino Unido para promover la seguridad en línea. Hay diferentes formas en que las empresas han intentado reducir los daños en línea, como tomar medidas una vez que reciben notificaciones de infracciones de políticas comunitarias, usar equipos de moderadores o utilizar medios tecnológicos para detectar y eliminar contenido perjudicial. Sin embargo, existen fuertes preocupaciones sobre la transparencia en la implementación de los términos y condiciones, y no hay coherencia, coordinación o estándares generales entre las plataformas en línea con respecto a contenidos y procedimientos no deseados (OHWP, pp. 34-38).

3. La distinción de dos niveles de daños en línea

Abordar los daños ilegales en línea ha sido una constante preocupación política en Europa y el Reino Unido. Un ejemplo de esto es la Estrategia del Mercado Único Digital para Europa (2015), que declaró la necesidad de evaluar el papel de los intermediarios y las plataformas en línea para crear un internet más seguro (COM/2015/0192, sección 3.3.2.). La participación de los intermediarios o los ISP ha sido especialmente discutible ya que "no son editores, pero tampoco son conductos neutrales; su papel en la gobernanza de los mercados de contenido en línea tiene inevitables connotaciones éticas" (Bunting, 2018, p. 165).

El OHWP planteó el mismo problema, pero añadiendo cierta controversia en relación con el concepto de daños en línea. Una primera distinción que introdujo fue la de contenido ilegal y contenido perjudicial. En varias ocasiones, el OWHP distingue claramente la conducta ilegal de las actividades perjudiciales (pp. 5, 11, 15, 42, 44, 66). Por obvio que parezca, los daños ilegales deben estar legalmente definidos o, al menos, la conducta ofensiva debe estar cubierta por una disposición legal para ser identificada como ilegal. Según la Comisión Europea, "contenido ilegal significa cualquier información que no cumple con el derecho de la Unión o el derecho de un Estado miembro afectado" (Comisión Europea, 2018, capítulo I, 4(b). Por otro lado, las actividades perjudiciales implican comportamientos que no son ilegales, pero se consideran inaceptables.

Ahora bien, a nivel europeo esta no es una distinción original del OHWP, ya que anteriormente, en la respuesta del gobierno al Libro Verde sobre la Estrategia de Seguridad en Internet (2008), se declaró expresamente que, aunque se habían logrado avances en la eliminación de material ilegal (especialmente relacionado con el terrorismo), se debían hacer mayores esfuerzos para reducir el contenido perjudicial, tanto legal como ilegal (OHWP, pp. 18-19.).

Esta clasificación de dos niveles de daños en línea (ilegales y legales pero perjudiciales) no es una distinción pacífica, ya que a veces los términos ilegales y perjudiciales parecen

utilizarse como términos equivalentes. Como ejemplo, la Comunicación final de la Comisión Europea sobre plataformas en línea y el mercado único digital, oportunidades y desafíos para Europa (COM/2016/288 final), destaca que

> existen áreas importantes como la incitación al terrorismo, el abuso sexual infantil y los discursos de odio en las que se debe alentar a todo tipo de plataformas en línea a tomar medidas voluntarias más efectivas para reducir la exposición a contenido ilegal o perjudicial. (p. 8.)

Resulta claro que en el contexto de la declaración, cuando la Comisión utiliza la conjunción "o" al referirse a contenido ilegal y perjudicial, está presumiendo que los términos son sinónimos, especialmente porque evoca términos ilegales como la incitación al terrorismo, el abuso sexual infantil y los discursos de odio. Sin embargo, el OHWP pareció alejarse de este enfoque, utilizando la conjunción "y" para destacar la distinción, lo que admitiría una clara diferencia entre esos términos. Esto plantea una distinción sutil pero relevante, junto con el problema de cómo diferenciar adecuadamente ambas categorías.

Habiendo dicho lo anterior, es importante señalar que los daños ilegales en línea no podrían reducirse a los enumerados como claramente definidos por el OHWP, ya que menciona expresamente que es una lista preliminar, no fija ni exhaustiva, basada en la predominancia que esos daños tienen en individuos y la sociedad (p. 30). Además, el OHWP excluyó de su alcance algunos daños en línea específicos (pp. 31-32)[5]. Sin embargo, parte de la doctrina insistió desde un primer momento que OHWP fue una colección "selectiva, inconsistente y jerárquica" de daños, especialmente porque no incluye los sufridos en línea por niñas y mujeres (Barker y Jurasz, 2019).

Además, resulta notorio que la regulación de los daños en línea es un área compleja para la política pública, pero al menos la regulación de los daños ilegales genera un consenso más amplio que la idea de regular contenido legal pero inaceptable[6]. El OHWP también dejó en claro que se centró en dos daños ilegales principalmente debido a su relevante impacto en individuos y la sociedad (la explotación y abuso sexual infantil y la actividad terrorista). En la doctrina de Reino Unido, hemos encontrado una tendencia similar de respaldo a acciones más estrictas y medidas de eliminación efectivas para delitos ilegales (relacionados con terrorismo, abuso infantil y discursos de odio) (Nash, 2019b, pp. 21-22). Sin embargo, el OHWP no estuvo exento de críticas respecto a los daños ilegales, pues se criticó la vaguedad de los daños preliminarmente enumerados como claramente definidos. Desde ese punto de vista, se afirmó que el OHWP oscureció intencionalmente la distinción para "ampliar su alcance mientras dificulta la identificación

[5] En esta línea, el OHWP excluye todos los perjuicios a las organizaciones (incluido el derecho de la competencia, la propiedad intelectual y la actividad fraudulenta), los perjuicios sufridos por los particulares en relación con las violaciones de datos y los perjuicios en línea sufridos por los particulares en la web oscura.

[6] Esta fue una de las conclusiones en "day-long multi-stakeholder workshop convened to discuss the implications of the 2019 Online Harms White Paper", resumido en Nash, 2019a, p. 3.

de las situaciones respecto de las que se regula como contenido legal" (Goldman, 2020, p. 359, y nota n° 34). Asimismo, se indicó que en realidad consistía en una mezcla de daños (Krasodomski-Jones, 2019)[7]. A modo de ilustración, y en relación con los daños enumerados como claramente definidos, no todos los tipos de acoso son ilegales y la definición de delito de odio es poco clara en el Reino Unido (p. 359).

En consecuencia, a pesar de que el OHWP declaró su intención de ser un nuevo marco regulatorio para mejorar la seguridad en línea en la economía digital, pareció ser una propuesta demasiado pretenciosa, ya que si bien pretendía regular todos los daños, solo esbozo algunos.

Con todo, dado lo problemático que resultó el primer listado de actividades ilegales abordadas por el OHWP, en la respuesta completa a la consulta de Diciembre 2020, el gobierno del Reino Unido señaló la necesidad de excluir de una nueva regulación sobre el contenido en línea, algunos de los daños que ya se encuentran cubiertos por otros cuerpos normativos, tales como aquellos relacionados con (i) la propiedad intelectual; (ii) protección de datos; (iii) fraude; (iv) protección de los consumidores; y (v) ciberseguridad o hacking[8].

Asimismo, en el mismo documento, el gobierno de Reino Unido definió que la legislación no exigiría la remoción de contenido legal pero potencialmente dañino. Sin embargo, se declaró que en una futura regulación se intentará introducir limitaciones a otros contenidos legales pero indeseados con el objeto de asegurar transparencia y uniformidad en el control de contenido en línea realizado por los ISP (tales como contenidos que promuevan la autolesión, contenidos que inciten al odio, abusos en línea que no alcancen el umbral de un delito penal y contenidos que fomenten o promuevan trastornos alimentarios)[9].

Finalmente, y a propósito de la discusión que generó la posibilidad de limitar la libertad de expresión a través de la censura previa del contenido legal pero perjudicial en línea, el Parlamento Británico decidió desistirse de insistir en este concepto. Dado lo anterior, el *Online Safety Act* se enfocó en la protección de los menores de edad (incluyendo contenido dañino para menos de edad) y la remoción únicamente de aquello identificado derechamente como ilegal (incluyendo contenido terrorista). Por lo tanto, luego de un amplio debate, la normativa no reguló el contenido legal pero dañino, concentrándose entonces en los contenidos primarios[10] y prioritarios[11] que son dañinos a los menos de edad.

[7] A mayor abundamiento, Smith (2019) ha señalado respecto al OHWP que "si el camino al infierno estaba pavimentado con buenas intenciones, esto era una autopista".

[8] Respuesta completa a la consulta sobre OWHP: https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response#part-3-the-regulator

[9] Ídem.

[10] En la Sección 7, el artículo 61 describe el contenido primario que puede resultar dañino para menores: (i) contenido pornográfico, (ii) contenidos que fomenten, promuevan o den instrucciones para el suicidio, (iii) contenido que fomente, promueva o proporcione instrucciones para un acto de autolesión deliberada, y (iv) contenidos que fomentan, promueven o proporcionan instrucciones para un trastorno alimentario o comportamientos asociados a un trastorno alimentario.

[11] En la Sección 7, el artículo 62 describe el contenido prioritario que puede resultar dañino para menores:

4. Examinando los daños legales pero perjudiciales en línea

Habiendo examinado la tensión que existe por distinguir los daños ilegales de los daños legales pero perjudiciales propuesto en el OHWP, es necesario abordar ahora el concepto más incierto y controvertido relacionado con el contenido legal pero perjudicial o nocivo. Recordemos que, bajo el término de contenido perjudicial (o daños con definición menos clara), el OHWP enumeró el ciberacoso y el *trolleo*, el contenido y la actividad extremistas, el comportamiento coercitivo, la intimidación, la desinformación, el contenido violento, la promoción de la autolesión y la promoción de la mutilación genital femenina (MGF).

La incertidumbre sobre este concepto viene dada porque el OHWP no definió el daño ni las actividades perjudiciales, ni describió elementos que contribuyeran a delimitar el término. Esto significa que la definición del término quedará en manos de la autoridad reguladora[12] (Smith, 2019b). La respuesta inicial al OHWP no aportó más aclaraciones sobre el tema (Smith, 2020). Sin embargo, cabe destacar que el problema asociado a un término deficientemente definido no es nuevo y se ha planteado hace bastante tiempo como un problema importante y no resuelto que debe abordarse cuidadosamente antes de imponer políticas regulatorias (ver Walker y Akdeniz, 1998, pp .9-10).

Además, el OHWP afirma expresamente que el enfoque regulatorio para el contenido legal pero perjudicial dependerá del contexto y que, eventualmente, esos daños incluirán riesgos relacionados con los daños emergentes de la tecnología que aún no han sido detectados (p. 42).

Debido al amplio abanico de daños y a la incertidumbre relacionada con el concepto de contenido perjudicial, se ha discutido que, basándose en las expresiones utilizadas por la OHWP, significaría cualquier cosa incivilizada, antidemocrática o perjudicial para la salud (Lesh et al., 2019)[13]. Estos conceptos, a su vez, implican un abanico de conductas

(i) contenidos que sean abusivos y que tengan impacto en la raza, religión, el sexo, la orientación sexual, discapacidad, o cambio de sexo, (ii) contenidos que inciten al odio en relación con las características antes descritas, (iii) contenidos que fomenten, promuevan o den instrucciones para cometer un acto de violencia grave contra una person; (iv) bulling; (v) contenido que represente violencia grave real o realista contra una persona o represente con detalle gráfico las lesiones graves reales o realistas de una persona; (vi) contenido que represente violencia grave real o realista o incluya un detalle gráfico de la lesión a un animal, incluyendo criaturas ficticias; (vii) contenidos que fomenten, promuevan o proporcionen instrucciones para un desafío o actividad que pueda provocar lesiones graves a la persona que lo realiza o a otra persona; (viii) contenidos que animen a una persona a ingerir, inyectar, inhalar o de cualquier otra forma autoadministrarse una sustancia físicamente nociva o una sustancia en una cantidad tal que sea físicamente nociva.

[12] En el caso del *Online Safety Act,* la autoridad reguladora del Reino Unido es OfCom.

[13] Donde se preguntan si "¿se nos permitirá criticar a los políticos con memes? ¿Constituyen 'desinformación' las fotos editadas con humor del autobús de campaña de Vote Leave? ¿Son las preguntas sobre la composición de la inmigración en el Reino Unido similares al 'contenido extremista'?". En sentido similar, Murray (2019, p. 5) indica el siguiente caso de difícil solución: "Pongamos un pequeño ejemplo: imaginemos que una plataforma de intercambio de vídeos que muestra las consecuencias de un ataque con drones en Siria. Es evidente que civiles inocentes han sido lastimados. El vídeo menciona la Yihad contra Estados Unidos, pero no menciona a ningún terrorista. ¿Se trata de un vídeo de actualidad o de noticias, o se trata de contenido terrorista? ¿Y si el vídeo mostrara las secuelas de un atentado en el Reino Unido y un comentarista mencionara la guerra contra el terrorismo? Este es sólo un ejemplo de una serie de juicios de valor muy

o actividades demasiado amplio como para ser previsto en una lista. Así pues, limitar los contenidos nocivos introduce un elemento ético que no está presente en el debate sobre los daños ilegales. Además, se ha criticado que es una ficción abordar los contenidos legales pero dañinos al tiempo que se resguarda la libertad de expresión, porque inevitablemente esta libertad será limitada o restringido durante la evaluación de la nocividad de la actividad (Wragg, 2019, p. 49). Además, se ha argumentado que donde la OHWP "fracasa más gravemente" es al tratar de regular de la misma manera contenidos dañinos pero legales con daños ilegales, ya que los primeros requieren un análisis más profundo (The Guardian, 2019).

Otros subrayaron que la intención de regular los daños más allá de los que prohíbe el derecho penal es una iniciativa alentadora, pero crea el riesgo de que dicha regulación se utilice de forma abusiva, influenciada por los medios de comunicación, el oportunismo político o por acontecimientos esporádicos que fomenten el descontento social ante situaciones concretas. Por lo tanto, puede ser muy subjetiva la delimitación de lo que es aceptable y, en consecuencia, se vuelve ambigua la manera de restringir la libertad de expresión en línea (Theil, 2019, p. 44). Si bien las políticas públicas basadas en daños pueden ser bienvenidas, requieren necesariamente de una base empírica que demuestre la gravedad del peligro que se pretende prevenir. En esta línea, la OHWP no expresó argumentos concretos a favor de una asociación efectiva entre el daño experimentado o causado con aquellas conductas catalogadas como inaceptables. Al basarse mayoritariamente en presunciones, la iniciativa fue criticada por carecer de justificación suficiente para limitar la libertad de expresión (Nash, 2019b, pp. 21-22). Del mismo modo, y debido a la falta de una base sólida de pruebas, es imposible predecir las consecuencias de los contenidos nocivos, ya que dependerán del contexto y de las percepciones personales de los usuarios (Nash, 2019a, p. 5).

A mayor abundamiento, se señaló que la forma en que la OHWP estableció la aplicación del deber de diligencia en relación con conductas legales pero dañinos puede suponer una infracción del "criterio de libertad de expresión del Convenio Europeo de Derechos Humanos, según el cual las restricciones deben estar prescritas por la ley y ser necesarias para un fin legítimo" (Tambini, 2019, p. 33). Por lo tanto, existe un riesgo de censura cuando el deber de diligencia se decide en códigos de conducta elaborados por las autoridades reguladoras, basándose en daños indefinidos o menos definidos. Las restricciones a la libertad de expresión deben pasar siempre por el debate legislativo y ser previstos en la legislación (p. 33).

Esto es especialmente atendido por Madiega (2020), al insistir que un punto de difícil solución es distinguir entre

complejos que el Gobierno pretende externalizar a las plataformas a través del deber de diligencia en línea".

> lo que es 'contenido ilegal en línea' de los contenidos que son 'nocivos' pero no ilegales, —pues— mientras que el concepto de 'nocivo' es subjetivo, depende en gran medida del contexto y puede variar considerablemente de un Estado miembro a otro. Además, los defensores de los derechos fundamentales sostienen que la introducción de normas para abordar los contenidos nocivos en línea en la legislación de la UE tendría graves consecuencias para la libertad de expresión, la libertad de buscar información y otros derechos fundamentales, por lo que pretenden limitar estrictamente el ámbito de aplicación de la ley de servicios digitales a los contenidos ilícitos. (p. 11)

Por último, cabe señalar que el enfoque de la OHWP fue más pretencioso que incluso las iniciativas nacionales sobre daños en línea, como la ley alemana de redes sociales (NetzDG), la ley francesa (Proposition de loi contre les contenus haineux sur Internet, 2020) (Cohen, 2019; Hoffmann y Gasparotti, 2020, pp. 25-26), y la Enmienda australiana del Código Penal para sancionar el intercambio de material violento desagradable (2019) (Murray, 2019, p. 3). Todas estas iniciativas difieren en elementos sustanciales, pero tienen en común que abordan principalmente los daños ilegales (no legales pero dañinos).

Un antecedente adicional de legislación comparada que ha intentado abordar el contenido dañino en línea es el proyecto presentado por el gobierno de Canadá con fecha 26 de febrero de 2024 (Bill C-63). En términos similares al Reino Unido, el propósito de este proyecto de ley es crear una Ley de Daños en Línea (Online Harms Act). Si bien por la denominación del proyecto pareciera que también habría intentado regular el contenido dañino pero legal, al revisar la propuesta, se observa que se priorizó regular siete tipos de daños en línea, lo que supone además modificar el Código Penal de Canadá. En ese sentido, se pretende impedir: (i) contenido que victimiza sexualmente a un niño o revictimiza a un superviviente; (ii) contenido íntimo compartido sin consentimiento; (iii) contenido utilizado para intimidar a un niño; (iv) contenido que induce a un niño a autoinfligirse daños; (v) contenido que fomente el odio; (vi) contenido que incita a la violencia; y (vii) contenidos que inciten al extremismo violento o al terrorismo[14].

Con todo cabe mencionar que el Reglamento de Servicios Digitales de la Unión Europea (EU Digital Services Act), que entró en vigencia en febrero de 2024, impone deberes de diligencia a las plataformas reguladas cuando se trate de actividades ilegales. En ese sentido, el considerando 12° de dicho cuerpo legal es claro al destacar que

> A fin de alcanzar el objetivo de garantizar un entorno en línea seguro, predecible y digno de confianza, para los efectos del presente Reglamento, el concepto de 'contenido ilícito' debe reflejar a grandes rasgos las normas vigentes en el entorno fuera de línea. Concretamente, el concepto de 'contenido ilícito' debe definirse de manera amplia para abarcar la información relacionada con contenidos, productos,

[14] El proyecto se encuentra disponible en https://www.parl.ca/LegisInfo/en/bill/44-1/c-63

> servicios y actividades de carácter ilícito. En particular, debe entenderse que dicho concepto se refiere a información, sea cual sea su forma, que sea de por sí ilícita en virtud del Derecho aplicable, como los delitos de incitación al odio o los contenidos terroristas y los contenidos discriminatorios ilícitos, o que las normas aplicables consideren ilícita por estar relacionada con actividades ilícitas.[15]

Es interesante mencionar que el Reglamento de Servicios Digitales impone a las plataformas y motores de búsqueda de muy gran tamaño adoptar medidas proporcionadas para evitar ciertos riesgos asociadas a conductas que podrían ser nocivas al discurso cívico y los procesos electorales, así como sobre la seguridad pública, la violencia de género, la protección de la salud pública y los menores y las consecuencias negativas graves para el bienestar físico y mental de la persona (artículo 34).

Lo anterior demuestra que el interés de regular el daño en línea encuentra cada vez más recepción en legislaciones comparadas. Si bien existe mayor consenso por avanzar con el control de conducta ilegales, se observa una tendencia por incentivar una co-regulación con las plataformas de mayor tamaño respecto de contenidos nocivos o perjudiciales a través de mecanismos de control diferenciados y proporcionales.

5. Contribución del discurso ilegal o indeseado como punto de partida para evaluar un concepto de daño en línea

Dado que la noción de contenido legal pero dañino carece de precisión, cualquier discurso o conducta de los usuarios podría eventualmente considerarse dañino dentro de la amplia gama de daños enumerados en la OHWP (Smith, 2019b). Sin embargo, con el objeto de delinear el concepto de daño legal pero perjudicial en línea, se ha precisado que aunque "el concepto de incitación al odio es uno de los más ampliamente debatidos y a la vez más esquivos en los estudios jurídicos" (Cavaliere, 2019, p. 6), parece que el estado actual del debate sobre el discurso ilegal o indeseado puede contribuir a desentrañar elementos relevantes para la noción de contenido perjudicial en línea (especialmente al evaluar el trolling, la intimidación, el contenido extremista e incluso la desinformación).

En esta línea, se ha planteado la existencia de una conexión entre cómo la regulación de contenidos nocivos pero legales en virtud de la OHWP conlleva en la práctica una regulación del discurso nocivo o indeseado (Haggart y Tusikov, 2019). Sin embargo, estimamos que el término "contenido nocivo en línea" es más amplio que el de "discurso nocivo" (ya que el primero circunscribe otras formas de daño distintas de las restringidas

[15] En ese sentido, el Reglamento de Servicios Digitales ejemplifica como conductas ilícitas el intercambio de imágenes que representen abusos sexuales de menores, el intercambio ilícito no consentido de imágenes privadas, el acoso en línea, la venta de productos no conformes o falsificados, la venta de productos o la prestación de servicios que infrinjan el Derecho en materia de protección de los consumidores, el uso no autorizado de material protegido por derechos de autor, la oferta ilegal de servicios de alojamiento o la venta ilegal de animales vivos (Considerando 12).

al segundo). Sin embargo, el debate sobre el discurso dañino parece adecuado para distinguir lo que es tolerable o no en el entorno digital.

Luego, en virtud de las características inherentes de internet y de los contenidos nocivos en línea, se hace necesario repensar, identificar e implementar nuevos criterios para fijar los umbrales de inaceptabilidad y también implementar medidas tecnológicas adecuadas para hacerlos cumplir; las cuales por defecto tienen que diferir de las tradicionalmente propuestas para otras áreas del derecho para lo considerado inaceptable.

5.1. Antecedentes de la distinción entre un discurso aceptable e inaceptable en Europa

La libertad de expresión ha estado especialmente protegida en Europa, teniendo su disposición más relevante en el artículo 10 del Convenio Europeo de Derechos Humanos[16].

Ahora bien, unos de los primeros documentos formales que menciona la distinción entre contenidos ilícitos y lícitos pero nocivos se remonta a 1996, cuando la Comisión de las Comunidades Europeas publicó una comunicación (COM/96/487 final) sobre contenidos ilícitos y nocivos en Internet, en la que se afirmaba que:

> en lo que se refiere a los contenidos ilícitos y nocivos, es fundamental diferenciar entre los contenidos que son ilícitos y otros contenidos nocivos. Estas diferentes categorías de contenidos plantean cuestiones de principio radicalmente distintas y exigen respuestas jurídicas y tecnológicas muy diferentes. (p. 10)

El documento enfatiza que los Estados miembros deben definir por ley las conductas ilegales y aplicarlas mediante la detección y el castigo de los infractores. Los contenidos nocivos, en cambio, consisten en materiales ofensivos para los sentimientos de las personas y deben tenerse en cuenta consideraciones culturales y éticas para establecer la frontera de lo que debe ser un material aceptable o permisible (COM/96/487 final, p. 11).

A nivel legislativo, la Directiva (UE) 2018/1808, que modificó la Directiva 2010/13/UE (Directiva de prestadores de servicios de comunicación audiovisual) establece nuevas normas para los servicios de plataformas de intercambio de vídeos y los servicios de medios sociales en relación con la protección de los niños y del público en general frente a los contenidos nocivos y la incitación al odio promovidos en línea en los servicios de plataformas de intercambio de vídeos (artículos 6 bis y 28 ter) (Montagnani y Trapova, 2019, pp. 6-7). También debe mencionarse que el Convenio sobre la Ciberdelincuencia (2001) define el material racista y xenófobo[17] y el Protocolo adicional al Convenio sobre

[16] Para referencias respecto a la regulación de la libertad de expresión en Europa ver McGonagle, 2020, pp. 474-480 y Iglezakis, 2017, pp. 367-283.

[17] El artículo 2 define el término como “cualquier material escrito, cualquier imagen o cualquier otra representación de ideas o teorías, que propugne, promueva o incite al odio, la discriminación o la violencia, contra cualquier individuo o grupo de individuos, por motivos de raza, color, ascendencia u origen nacional o étnico, así como la religión si se utiliza como pretexto para cualquiera de estos factores”.

la Ciberdelincuencia (2003) exige a los Estados miembros que introduzcan sanciones penales por comportamientos racistas y xenófobos o por negación o justificación del genocidio o de crímenes contra la humanidad (artículos 3 a 6).

Desde una perspectiva voluntaria, el Código de Conducta de la Unión Europea para combatir la incitación ilegal al odio en línea (Código de Conducta) fue acordado entre la Comisión y algunas de las principales plataformas de medios sociales[18]. El Código de Conducta se basa en la definición de incitación ilegal al odio recogida en la Decisión Marco 2008/913/JAI, de 28 de noviembre de 2008, que se centra en delitos relacionados con (i) racismo y xenofobia relacionados con la difusión pública y la incitación a la violencia o al odio dirigidos contra un grupo de personas o un miembro de dicho grupo y relativos a la raza, el color, la religión, la ascendencia o el origen nacional o étnico; (ii) negar públicamente el genocidio, los crímenes contra la humanidad y los crímenes de guerra; y (iii) negar de manera que pueda incitar a la violencia o al odio contra un grupo específico de personas basado en el racismo o la xenofobia (artículo 1 n. ° 1).

El Código de Conducta reconoce que la incitación ilegal al odio fuera de línea se aborda mediante un sólido sistema de aplicación de la ley penal, pero la incitación ilegal al odio en línea requiere directrices para que los intermediarios en línea actúen con prontitud ante esos daños. Por lo tanto, las empresas en línea están obligadas a establecer sus políticas comunitarias para cumplir con las normas establecidas por el Código de Conducta.

Incluso con la protección legal contra el discurso de odio, la distinción entre discurso apropiado o inaceptable parece ser un área recurrente y gris de debate. Se pueden encontrar referencias a esta distinción en la Comunicación de la Comisión Europea sobre la lucha contra los contenidos ilícitos en línea (COM/2017/0555), cuyo objetivo es ofrecer directrices a las plataformas en línea sobre cómo detectar y eliminar eficazmente los contenidos ilícitos[19]. Los académicos han encontrado declaraciones en el documento que sugerirían "dos categorías distintas de discurso 'malo': contenido ilegal, definido por las leyes nacionales y de la UE, y contenido indeseable definido por las propias plataformas" (Cavaliere, 2019, p. 5).

En ese sentido, la Comunicación sobre la lucha contra los contenidos ilícitos en línea afirma que

> las plataformas en línea deben ofrecer una explicación clara, fácilmente comprensible y suficientemente detallada de su política de contenidos en sus condiciones de servicio. Esto debería reflejar tanto el tratamiento de los contenidos ilegales como el de los contenidos que no respetan las condiciones de servicio de la plataforma.

Además,

[18] Los primeros firmantes fueron Facebook, Microsoft, Twitter, YouTube y servicios de consumo alojados en Microsoft, como los servicios de juegos de Xbox o LinkedIn, a los que se unieron después Instagram, Google+, Snapchat, Dailymotion y Jeuxvideo.com.

[19] Como el terrorismo, la incitación ilegal al odio, el abuso de menores o la trata de seres humanos

> la cuestión de si un contenido es legal o ilegal se rige por las leyes nacionales y de la UE. Al mismo tiempo, las propias condiciones de servicio de las plataformas en línea pueden considerar indeseables o censurables determinados tipos de contenidos. (Cavaliere, 2019, p. 16)[20]

A las afirmaciones anteriores, añadiríamos las siguientes del mismo documento:

> No cabe duda de que existen preocupaciones de interés público en torno a contenidos que no son necesariamente ilegales pero potencialmente nocivos, como las noticias falsas o los contenidos perjudiciales para los menores. Sin embargo, la presente Comunicación se centra en la detección y eliminación de contenidos ilegales. (COM/2017/0555, p. 6).

Además, el documento menciona que el Parlamento Europeo en 2017 introdujo esta distinción ya que se recomendó a las plataformas "reforzar las medidas para hacer frente a los contenidos ilícitos y nocivos" (Cavaliere, p. 2)[21].

La distinción también se puede encontrar en las Directrices de Derechos Humanos para los ISP, desarrolladas por el Consejo de Europa en cooperación con la Asociación Europea de Proveedores de Servicios de Internet, donde los ISP pueden encontrar orientación sobre cómo abordar los contenidos ilícitos y nocivos (este último especialmente centrado en la protección de los niños) (Consejo de Europa y EuroISPA, 2008).

Todas las referencias revisadas muestran que existe una tendencia política consistente en diferenciar los contenidos ilegales de los dañinos y, en relación con estos últimos, la necesidad de distinguir entre lo que es un discurso legal no deseado o aceptable. Sin embargo, no se advierte claridad al momento de identificar el discurso legal no deseado, permaneciendo en un área gris de difícil solución.

5.2. Elementos para evaluar la adecuación del discurso

El discurso de odio no abarca todo tipo de discurso. Normalmente, el discurso de odio está relacionado con la intolerancia extrema dirigida a un grupo específico. El discurso intolerante no cumple el estándar extremo o ultrajante del discurso de odio, por lo que no debería estar prohibido por la ley. De hecho, la antipatía, el desacuerdo o la intolerancia forman parte de la naturaleza humana, que en algunos casos puede ser incluso necesaria o positiva, como cuando se trata de debatir sobre la corrupción o la injusticia (Post, 2009, pp. 123 y 125).

[20] El autor también encuentra antecedentes de la distinción en Comisión Europea, Recommendation of 1.3.2018 on measures to effectively tackle illegal content online C(2018)1177 final, considerando 23. http://data.europa.eu/eli/reco/2018/334/oj

[21] Haciendo alusión a la Resolución del Parlamento Europeo de 15 de junio 2017 respecto a plataformas en línea (2016/2274/INI).

Sin embargo, toda regulación del discurso es especialmente compleja pues supone, en algún grado, soslayar al derecho a emitir opinión. En ese sentido, Riso advierte que

> sin duda la regulación del discurso de odio, así como toda limitación de la libertad de la comunicación de pensamiento, implica ingresar en una pendiente resbaladiza en la que es muy difícil mantener el equilibrio. Más cuando el clamor popular es a favor de una causa noble: terminar con la exclusión, proteger a sujetos que están siendo dañados y discriminados, etc. (Risso Ferrand, 2020, p. 74)

Ahora bien, Post sostiene que la incitación al odio debe evaluarse no sólo en relación con el contenido del discurso, sino con la forma en que se presenta. Esto se refiere al estilo en que se estructura el insulto, la ofensa o la degradación. El discurso sobre la raza, la nacionalidad, la sexualidad o la religión puede parecer decente y aceptable, pero el umbral que establece la distinción entre lo que constituye discurso de odio debe medirse en función de las normas sociales. En relación con la norma social el autor aclara que

> utilizaré el término 'normas' para referirme a las actitudes grupales que todos llevamos dentro todo el tiempo y desde el fundamento y la posibilidad de nuestro propio 'yo', y utilizaré el término 'comunidad' para referirme a la forma de organización social que se crea y se sostiene en dichas normas. (Post, 2009, p. 128)

Del mismo modo, se ha argumentado que, dado que los ISP tienen un gran poder en relación con la interacción en línea de las personas, deben asumir un rol y una responsabilidad cívica en relación con la forma en que operan su negocio. Según este punto de vista,

> en los contextos internacionales y multiculturales en los que operan los ISP, la especificación de lo que es socialmente aceptable y preferible sólo será efectiva —es decir, se considerará éticamente sólida, apropiada y deseable— en la medida en que se apoye en un enfoque capaz de conciliar los diferentes puntos de vista éticos y los intereses de las partes interesadas a los que se enfrentan los ISP. (Taddeo, 2020, p. 136)[22]

Adicionalmente, se han identificado cuatro variables subsumidas en las diferentes regulaciones del discurso, siendo cada una de ellas discutible, especialmente cuando se trata de la forma en que los ISP las han implementado en sus normas comunitarias (redes sociales). La primera tiene que ver con el ámbito de protección y el grupo de personas amparadas por las disposiciones, las cuales pueden estar basadas en religión, orientación sexual, discapacidades, raza, entre otras. La segunda variable tiene que ver con la forma del discurso, en cuanto a su capacidad de promover conductas o acciones

[22] En el mismo sentido, Helberger et al., 2018, p. 7.

nocivas (incluyendo representaciones escritas, gráficas o en video). La tercera variable tiene que ver con la naturaleza del daño, refiriéndose a la relación entre daño físico y no físico como objetos de regulación. Mientras que el daño físico ha sido el tradicional objeto de regulación, el no físico requiere una prueba de razonabilidad. La última variable tiene que ver con el vínculo causal entre el discurso y el daño (Cavaliere, 2019, pp. 8-27).

Estas serán las variables que utilizaremos a continuación para delinear y perfilar un concepto de daño en línea.

6. Categorización de los contenidos legales y perjudiciales a partir de los elementos del discurso de odio

La complejidad de la definición de contenidos nocivos en línea reside en el hecho de que implica diversos conceptos que difieren sustancialmente entre sí, especialmente en la forma en que ha sido descrito por la OHWP. Como se adelantó, para abordar este concepto indefinido, se analizará bajo la taxonomía de daños del OHWP que se encuentra en los grupos identificados como "daños con una definición menos clara" y "exposición de menores a contenidos legales"[23].

6.1. Primera clasificación: Daños individuales y sociales

Además de la distinción original de los daños en línea (daños ilegales o legales), la conducta dañina puede clasificarse utilizando la distinción entre daños sociales y daños individuales. Se ha dicho que los daños sociales pueden definirse claramente y su contenido responde a una naturaleza objetiva. En cambio, los daños individuales, no son fácilmente reducibles a un concepto definido y responden a hechos concretos. Su valoración requiere la ponderación de elementos subjetivos para establecer normas reguladoras (Tomlinson, (2019).

Esta distinción parece compatible con la OHWP, cuando menciona que los daños en línea "socavan nuestros valores y principios democráticos" (p. 5), y bajo el subtítulo "Amenazas a nuestro modo de vida", el documento describe cómo la desinformación y la difusión de información inexacta o falsa pueden ser perjudiciales. Pero al revisar todos los contenidos nocivos enumerados por el OHWP, hay una clara tendencia a centrarse en los daños individuales más que en los daños sociales. Por eso, de los contenidos enumerados en el OHWP, sólo la desinformación puede ajustarse a esta clasificación, dejando abierto el problema de categorizar para el resto de los daños individuales.

[23] Se hace presente que existen otras taxonomías propuestas de daños en línea como aquellos relacionados con los ciberdaños pero ajenas al OHWP como se examina en Agrafiotis, et al., 2018, p. 8. También el regulador del Reino Único, OfCom, ha proporcionado su propia taxonomía que se puede revisar en OfCom, 2018, p. 12.

6.2. Segunda clasificación: Perjuicios que afectan indistintamente a todos los individuos y los que afectan a un grupo especial de personas en función del sexo o la edad

Los daños individuales admiten una nueva subclasificación entre los contenidos inaceptables dirigidos a un grupo especial de personas en función de su género o edad y aquellos daños que afectan indistintamente a todos los individuos.

De la lista de daños de la OHWP, la Mutilación Genital Femenina (MGF) puede clasificarse dentro de los contenidos inaceptables por razón de género, pero también cualquier otro contenido nocivo no regulado e imprevisto, como los contenidos sexualizados y/o misóginos (Barker y Jurasz, 2019), la discriminación salarial injustificada o el fomento de los trastornos alimentarios. Cabe señalar que estos contenidos también pueden afectar a hombres. Por otro lado, los niños que acceden a pornografía y los niños que acceden a material inapropiado pueden clasificarse como contenidos inaceptables para su edad. Este subgrupo también podría incluir contenidos considerados inadecuados para las personas mayores (ya que podrían ser un grupo vulnerable en el entorno digital)[24].

Nuestra decisión de utilizar esta subclasificación y no la variable relativa a los grupos de personas que se encuentran en el discurso del odio (basado en la religión, la orientación sexual, las discapacidades, la raza u otro elemento distintivo) se debe a que todos ellos deberían incluirse en el término "discurso del odio", que ya figura como daño ilegal. Sin embargo, esto abre la cuestión de cómo categorizar el discurso legal pero inaceptable en el OHWP. Creemos que esto puede resolverse utilizando dos estrategias diferentes: incluyéndolo como un nuevo concepto en la lista de "términos menos definidos" o subsumiéndolo en un término ya existente. Preferimos esta última, subsumiéndolo en "contenido violento", sobre todo porque la OHWP declara que "el contenido que es violento con una comprensión contextual adicional" (p. 67)[25] puede subsumirse en el concepto de "contenido violento". Consideramos que éste sería el caso de un discurso legal pero inaceptable.

6.3. Tercera clasificación: daños físicos y psicológicos

Por último, un elemento adicional del discurso que puede ser útil para desarrollar una última subclasificación es la naturaleza del daño, pues se basa en el impacto físico o psicológico sobre la víctima. El OWHP sólo se refiere a las consecuencias físicas del daño cuando se trata de terrorismo o de niños (pp. 41-43), pero no necesariamente para

[24] En términos generales el Servicio Nacional del Consumidor ha identificado a ciertos grupos de personas como consumidores vulnerables a través de la Circular Interpretativa sobre la Noción de Consumidor Hipervulnerable de fecha 31 de diciembre de 2021. Lo anterior ha sido comentado por López, 2022, pp. 340-415.

[25] En esta línea, el OHWP afirma "Los contenidos violentos van desde los que muestran directamente actos de violencia o incitan a ellos, hasta los que son violentos con una comprensión contextual adicional o que son perjudiciales para los usuarios por la glorificación de las armas y la vida en pandilla".

otros tipos de daños (por ejemplo, los que pueden afectar a adultos, ancianos o personas con discapacidad).

Partiendo de la base de que la mayoría de los daños se han enumerado como "con una definición menos clara", la fragmentación propuesta puede variar en función del significado que cada sociedad atribuya a cada uno de los daños.

Dicho esto, la apología a la autolesión y los contenidos violentos pueden incluirse en los daños físicos. Por otra parte, el resto de los daños no categorizados enumerados en el OHWP (ciberacoso y trolling, contenido y actividad extremista, comportamiento coercitivo e intimidación) pueden clasificarse dentro de los daños psicológicos.

7. Propuesta de un concepto de daño en línea en virtud de la categorización de los daños propuesta

Como se ha señalado, el problema de conceptualizar el término "contenido dañino en línea" surge del hecho que su definición no se encuentra en el OHWP o en otros textos legales, y que a su vez, refiere a otros conceptos indefinidos. Dado lo anterior, resulta de mayor complejidad definir un subgrupo de daños (contenido legal pero perjudicial). En otras palabras, si el género (daños en línea) carece de definición, más difícil resulta definir la especie (contenido legal pero perjudicial) y, en consecuencia, categorizar cada una de las subespecies (cada contenido nocivo).

El problema de las definiciones es que tienden a ser conceptos cerrados, carentes de la flexibilidad necesaria para adaptarse a la realidad que intentan describir. Alcanzar el equilibrio correcto entre certidumbre y flexibilidad es un esfuerzo difícil pero necesario para ofrecer una definición clara, útil y duradera, especialmente para las ISP que necesitan cumplir con un deber de diligencia respecto a dichas conductas. En ese sentido, los agentes interesados o sujetos de posibles deberes de diligencia identificaron desde un comienzo que resultaba esencial que la regulación se impusiera respecto de conceptos claramente definidos (Center for Democracy & Technology, 2019, p. 2).

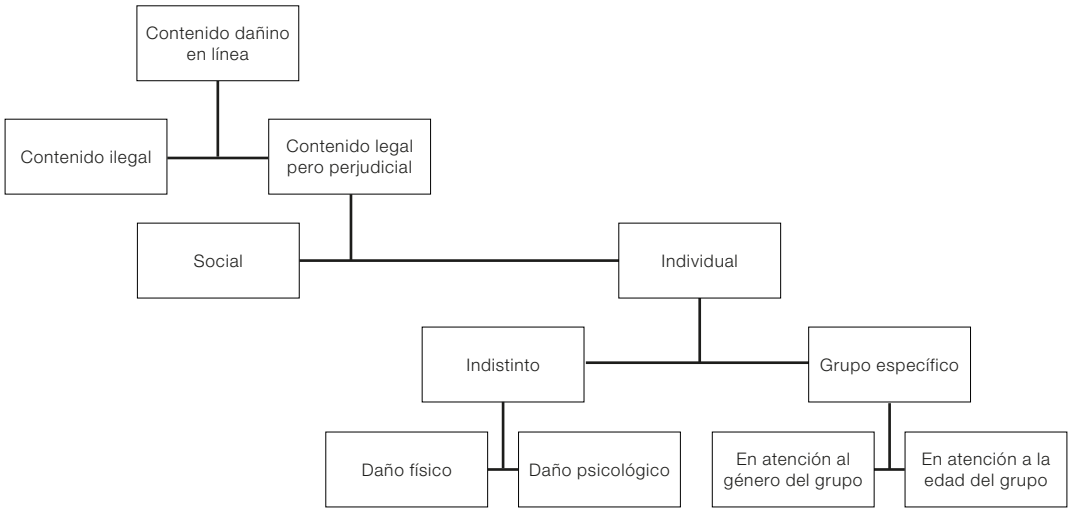
Luego, una definición demasiado concisa puede excluir nuevos elementos (conductas) que no estaban previstas al construir el término. Por otro lado, un término demasiado amplio no proporcionará certeza suficiente de lo que realmente significa y se espera de aquellos que deben velar por su cumplimiento.

Como ejemplo de este último enfoque, el "daño en línea" (que es un término más amplio que el de contenido legal pero dañino en línea) se ha definido de forma poco clara como cualquier "comportamiento en línea que pueda herir física o emocionalmente a una persona. Puede tratarse de información perjudicial publicada en línea o enviada a una persona" (OHWP). Del mismo modo, se ha subrayado que el contenido legal pero perjudicial "se refiere al contenido que a menudo no entra estrictamente en la prohibición de una ley, pero que sin embargo podría tener efectos nocivos" (Madiega, 2020, p. 10). Estas definiciones, aparte de la distinción de si el contenido entra o no en el ámbito de aplicación de la ley, no aportan más detalles ni claridad a los términos.

Por ello, la categorización de los diversos elementos circunscritos en el término “contenidos legales pero dañinos en línea” constituye un esfuerzo necesario para proponer un concepto más objetivo, ya que ayuda a comprender el objeto de estudio, al tiempo que proporciona un enfoque más detallado de cada subgrupo de conductas indeseadas.

A modo de recapitulación, el siguiente cuadro (ilustración 1) resume la categorización propuesta en el presente trabajo de los daños en línea:

Ilustración 1
Propuesta para delimitar elconcepto de “contenido dañino el línea”



En nuestra opinión, y en consideración a los elementos anteriormente categorizados, el concepto de “contenido dañino en línea” puede describirse como:

> Una conducta realizada en el entorno digital, expresada en cualquier forma o conjunto de representaciones, que aun siendo legal puede ser potencialmente perjudicial para una comunidad, un grupo específico de personas o para todo individuo, sobre la base de un entendimiento compartido y común de lo que se considera inaceptable en una sociedad democrática específica.

De la noción propuesta pueden destacarse cinco elementos clave:

En primer lugar, diferencia inmediatamente los contenidos legales y perjudiciales de los daños ilegales.

En segundo lugar, reconoce que cualquier representación o conjunto de ellas (imágenes, mensajes de texto, vídeos o audio) puede expresar un contenido nocivo, introduciendo un elemento contextual para la evaluación. Esto, a su vez, introduce flexibilidad en la definición.

En tercer lugar, distingue entre daños sociales (cuando se refiere a la comunidad) e individuales.

En cuarto lugar, utiliza subgrupos genéricos (afectan indistintamente a individuos o sólo a grupos específicos), en lugar de intentar identificar daños específicos. Esto también otorga adaptabilidad a la definición, ya que los daños imprevistos pueden atribuirse a esos grupos o a otros nuevos.

En quinto lugar, los motivos que marcarán el umbral entre lo aceptable y lo inaceptable no pueden ser arbitrarios, y por eso deben basarse en un debate que represente un acuerdo democrático en una sociedad concreta (lo que introduce la noción de normas sociales propuesta por Post en relación con el discurso legal pero perjudicial[26]). Como se trata de una distinción sutil, lo más probable es que la valoración varíe de un país a otro. Sin embargo, la delimitación de los objetos de regulación (contenidos inaceptables identificados) garantizará la protección de la libertad de expresión e información en cada sociedad y, en consecuencia, permitirá una regulación más objetiva respecto de la responsabilidad que se le exija a los ISP para controlar los contenidos indeseados.

8. Conclusiones

Es indesmentible que internet ha contribuido al desarrollo y avance de las sociedades. Sin embargo, internet también puede ser fuente de graves daños para las personas. Aquellas conductas realizadas en línea pueden afectar los derechos y libertades de los usuarios, al menos en el mismo grado como si se realizaran en el mundo material.

Con todo, determinar qué debe entenderse por daño en línea no se ha abordado con detención en la dogmática nacional y continúa siendo un concepto indeterminado. Lo anterior no es baladí pues un correcto entendimiento de dicha noción servirá para la construcción de políticas públicas adecuadas para prevenir y controlar aquel contenido que se estima inadecuado o intolerable. Asimismo, de esta definición se podrán articular sistemas regulatorios coherentes y orgánicos respecto del amplio espectro de conductas perjudiciales (actualmente descritas y nuevas que se originen en el futuro), y de esa manera, se evitarán respuestas fragmentadas en el ordenamiento nacional, como ha sido el caso de Chile.

En este trabajo se ha utilizado el debate generado a raíz del OHWP del Reino Unido para proponer un concepto de daño en línea, pues fue un documento que innovó al intentar introducir una distinción de dos niveles respecto de los daños en línea, esto es, los daños ilegales y los daños legales pero perjudiciales.

[26] Las cuestiones subjetivas, contextuales y domésticas relativas al concepto legal pero perjudicial han sido claramente abordadas por Madiega, 2020, p. 11.

Si bien la regulación de los daños en línea es un área de constante debate, al menos los daños ilegales admiten un mayor grado de conceso para su regulación. Lo anterior queda refrendado en las leyes de control de contenido en línea de Alemania, Francia, Australia, Reino Unido, la Unión Europea y el reciente proyecto de ley de Canadá que están orientados preferentemente a conductas ilegales (vinculadas con el terrorismo, la protección de niños y el discurso de odio).

Al contrario, el concepto de daño legal pero perjudicial introducido por el OHWP resulta ser más complejo y controvertido pues supone conductas vagamente definidas y que no es posible radicar en un cúmulo determinado de acciones. El riesgo de regular estas conductas es otorgar un grado de discrecionalidad mayor a los ISP al momento de controlarlos, lo que puede conllevar riesgos para la libertad de expresión de los usuarios al ser objeto de medidas de censura previa. Por lo tanto, existe un debate pendiente respecto a la tolerabilidad de ciertas conductas, que sin ser ilegales per se, de todas maneras se considerarán indeseadas. Para efectos metodológicos, se ha utilizado la taxonomía ofrecida por el OWHP respecto a este tipo de daños para delimitar las conductas analizadas.

En este artículo se ha propuesto que es posible perfilar el daño legal pero perjudicial utilizando el debate existente respecto al discurso nocivo, en donde existe evidencia a nivel comparado respecto a la necesidad de distinguir entre el discurso legal pero inaceptable. En ese sentido, hemos propuesto que las distintas conductas que podrían circunscribirse en el daño legal pero perjudicial pueden categorizarse utilizando la distinción entre daños individuales y sociales. Luego, dentro de los daños individuales, es necesario distinguir entre aquellas conductas que afectan indistintamente a todos los individuos de una sociedad (admitiendo un subgrupo de daños físicos y psicológicos), y aquellas conductas orientadas a grupos específicos (sea en función de su género o edad).

Esta clasificación es relevante para adoptar regulaciones específicas respecto a cada una de las categorías de daños. Asimismo, su categorización es un paso previo para ofrecer una definición de daños en línea, toda vez que las conductas aisladamente consideradas carecen de un orden que permita su conceptualización.

La anterior proposición ha permitido radicar las distintas conductas previstas por el OHWP en una categorización clara. En ese sentido, la primera observación a los daños en línea es la de realizar una distinción de dos niveles, consistente entre daños ilegales y daños legales pero perjudiciales.

Luego, dentro de los contenidos ilegales deben circunscribirse todas aquellas conductas claramente definidas en el ordenamiento y que tienen una sanción específica. Ahora bien, respecto de los contenidos legales pero perjudiciales, la desinformación estaría comprendida dentro de los daños que afectan a la comunidad o sociedad en su conjunto. A nivel individual, la apología de la autolesión y los contenidos violentos son atribuibles a posibles daños físicos en los individuos. Al contrario, las conductas de ciberacoso, trolling, contenido y actividad extremista, comportamiento coercitivo e intimidación descritas

en el OHWP quedarían comprendidas en los daños psicológicos que afectan a los individuos. Finalmente, conductas que afectan a un grupo en función de su género serían el fomento de la Mutilación Genital Femenina (MGF), contenidos sexualizados, misóginos, discriminación salarial injustificada o los trastornos alimentarios. A su vez, conductas que afectan a un grupo en función de su edad serían el acceso a pornografía por parte de niños y el acceso a otro tipo de materiales inapropiados para niños.

En base a lo anterior, y recogiendo la distinción de dos fases, así como los elementos que se han aportado para distinguir el discurso tolerable del intolerable, este trabajo propone definir el daño en daño en línea como la conducta realizada en el entorno digital, expresada en cualquier forma o conjunto de representaciones, que aun siendo legal puede ser potencialmente perjudicial para una comunidad, un grupo específico de personas o para todo individuo, sobre la base de un entendimiento compartido y común de lo que se considera inaceptable en una sociedad democrática específica.

Acerca del artículo

Notas de conflicto de interés. El autor declara no tener ningún conflicto de interés en relación con la publicación de este artículo.

Contribución en el trabajo. El autor asumió todos los roles establecidos en Contributor Roles Taxonomy (CRediT).

Referencias

Agrafiotis, I., Nurce, J., Goldsmith, M., CReese, S. y Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity* 4(1), tyy006. https://doi.org/10.1093/cybsec/tyy006

Barker, K. y Jurasz, O. (2019). ¿Online harms and Caroline's Law – what's the direction for the law reform? Scripted. https://script-ed.org/blog/online-harms-and-carolines-law/

Bunting, M. (2018). From editorial obligation to procedural accountability: policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy* 3(2). https://doi.org/10.1080/23738871.2018.1519030

Cavaliere, P. (2019). Digital platforms and the rise of global regulation of hate speech. *Edinburgh School of Law Research Paper* (2019/29). https://dx.doi.org/10.2139/ssrn.3456141

Center for Democracy & Technology (2019). *Nine Principles for Future EU Policymaking on Intermediary Liability*. CDT. https://cdt.org/wp-content/uploads/2019/08/Nine-Principles-for-Future-EU-Policymaking-on-Intermediary-Liability-Aug-2019.pdf

Cohen, M. (2019). Will the Online Harms White Paper make the UK the safest place in the world to go online? A look at recent approaches the UK, Germany, Australia and New Zealand have taken to regulating online harms. *Computer and Telecommunications Law Review*, 25.

Consejo de Europa y EuroISPA. (2008). *Human Rights Guidelines for Internet Service Providers*. COE. https://rm.coe.int/16805a39d5

Goldman, E. (2020). The U.K. Online Harms White Paper and the Internet's Cable-ized Future'. *Ohio State Tech.* L.J. 16(2). https://dx.doi.org/10.2139/ssrn.3438530

Haggart, R. y Tusikov N. (2019). What the UK's Online Harms white paper teaches us about internet regulation. *Inforrm*. https://inforrm.org/2019/04/22/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-richard-haggart-and-natasha-tusikov/

Helberger, N., Pierson J. y Poell, T. (2018), Governing online platforms: From contested to cooperative responsibility, *The Information Society*, 34(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913

Hoffmann, A. y Gasparotti, A. (2020). *Liability for illegal content online Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a Digital Services Act.* CepStudy. https://www.cep.eu/fileadmin/user_upload/hayek-stiftung.de/cepStudy_Liability_for_illegal_content_online.pdf

Iglezakis, I. (2017). The Legal Regulation of Hate Speech on the Internet. En T.-E. Synodinou, P. Jougleux, P., Markou, C. y Prastitou, T. (Eds.), *EU Internet Law. Regulation and Enforcement*. Sringer.

Krasodomski-Jones, Alex (2019). Can the government nudge us towards a better internet? CapX. https://capx.co/can-the-government-nudge-us-closer-to-a-better-internet/

Lesh, M., Dumitriu, S. y Salter, P. (2019). Safeguarding progress: The risks of internet regulation. Adam Smith Institute. https://coilink.org/20.500.12592/ck82jd

López Díaz, P. (2022). El consumidor hipervulnerable como débil jurídico en el derecho chileno: una taxonomía y alcance de la tutela aplicable. *Latin American Legal Studies* 10(2), 340-415. https://doi.org/10.15691/0719-9112Vol10n2a7

Madiega, T. (2020). Reform of the EU liability regime for online intermediaries. Background on the forthcoming digital service act. European Parliament Research Service. EPRS. https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf

McGonagle, T. (2020). Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation. En G. Frosio (Ed.), *Oxford Handbook of Online Intermediary Liability.* https://doi.org/10.1093/oxfordhb/9780198837138.013.24

Montagnani, M. y Trapova A. (2019). New Obligations for Internet Intermediaries in the Digital Single Market - Safe Harbors in Turmoil? *Journal of Internet Law*, 22(7), 3-11. https://dx.doi.org/10.2139/ssrn.3361073

Murray, A. (2019). Rethinking Regulation for the Digital Environment'. *LSE Law - Policy Briefing Paper*, (41). https://dx.doi.org/10.2139/ssrn.3462792

Nash, V. (2019a). *Internet Regulation and the Online Harms White Paper Stakeholder Workshop*. https://dx.doi.org/10.2139/ssrn.3412790

Nash, V. (2019b). Revise and resubmit? Reviewing the '2019 Online Harms White Paper' *Journal of Media Law*, 11(1), 18-27. https://doi.org/10.1080/17577632.2019.1666475

OfCom (2018). *Addressing harmful online content: A perspective from broadcasting and on-demand standards regulation.* https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/addressing-harmful-online-content/

Post, R. (2009). Hate Speech. En I. Hare y J. Weinstein (Eds.), *Extreme speech and democracy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548781.003.0008

Prince Torres, Á. C. (2020). El acceso a Internet como derecho fundamental: perspectivas internacionales. *Revista Justicia & Derecho*, 3(1), 1-19. https://doi.org/10.32457/rjyd.v3i1.45

Risso Ferrand, M. (2020). La libertad de expresión y el combate al discurso del odio. *Estudios Constitucionales*, 18(1), 51-89. http://dx.doi.org/10.4067/S0718-52002020000100051

Smith, G. (2019a, 30 de abril). Users Behaving Badly: the Online Harms White Paper. *Inforrm*. https://inforrm.org/2019/04/30/users-behaving-badly-the-online-harms-white-paper-graham-smith

Smith, G. (2019b, 12 de mayo). The Rule of Law and the Online Harms White Paper. *Cyberleagle*. https://www.cyberleagle.com/2019/05/the-rule-of-law-and-online-harms-white.html

Smith, G. (2020, 20 de febrero). Online Harms Deconstructed: the Initial Consultation Response. *Inforrm*. https://inforrm.org/2020/02/20/online-harms-deconstructed-the-initial-consultation-response-graham-smith

Taddeo, M. (2020). The Civic Role of OSPs in Mature Information Societies. En G. Frosio (Ed.), *Oxford Handbook of Online Intermediary Liability*. https://dx.doi.org/10.2139/ssrn.3584187

Tambini, D. (2019). The differentiated duty of care: a response to the Online Harms White Paper. *Journal of Media Law*, 11(1), 28-40. https://doi.org/10.1080/17577632.2019.1666488

The Guardian view on online harms: white paper, grey areas. *The Guardian*. https://www.theguardian.com/commentisfree/2019/apr/08/the-guardian-view-on-online-harms-white-paper-grey-areas

Theil, S. (2019). The Online Harms White Paper: comparing the UK and German approaches to regulation. *Journal of Media Law*, 11(1), 41–51. https://doi.org/10.1080/17577632.2019.1666476

Tomlinson, H. (2019). Online Harms White Paper: Two comments on harms. *Inforrm*. https://inforrm.org/2019/07/05/online-harms-white-paper-two-comments-on-harms-hugh-tomlinson-qc/

Walker, C. y Akdeniz, Y. (1998). The governance of the Internet in Europe with special reference to illegal and harmful content. *Criminal Law Review*, 5-19.

Wragg, P. (2019). Tackling online harms: what good is regulation? *Communications Law 24*(2), 49-51.

Normas citadas

Normativa chilena

Decreto 12. (12 de mayo de 2023). Crea comisión asesora ministerial del ministerio de ciencia, tecnología, conocimiento e innovación, denominada "comisión asesora contra la desinformación" [Ministerio de Ciencia, Tecnología, Conocimiento e Información]. [DEROGADO] https://bcn.cl/3e7zs

Decreto 5. (6 de marzo de 2024). Pone término a "comisión asesora contra la desinformación" [Ministerio de Ciencia, Tecnología, Conocimiento e Información]. https://bcn.cl/3khgd

Ley 18314. (17 de mayo de 1984). Determina conductas terroristas y fija su penalidad. https://bcn.cl/2k8cn

Ley 19733. (04 de junio de 2001). Sobre libertades de opinión e información y ejercicio del periodismo. https://bcn.cl/2f8z0

Ley 21522. (30 diciembre de 2022) introduce un nuevo párrafo en el título VII del libro II del código penal, relativo a la explotación sexual comercial y material pornográfico de niños, niñas y adolescentes. https://bcn.cl/3ayhg

Normativa europea

Comunicación COM/96/487 final de la Comisión Europea. (16 de octubre de 1996). Illegal and harmful content on the internet. https://eur-lex.europa.eu/procedure/EN/20878

Comunicación COM/2015/0192 final de la Comisión Europea (6 de mayo de 2015). A Digital Single Market Strategy for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52015DC0192

Comunicación COM/2016/288 final de la Comisión Europea. (25 de mayo de 2016). Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016DC0288

Comunicación COM/2017/0555 final de la Comisión Europea. (28 de septiembre de 2017). Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017DC0555

Directiva (UE) 2018/1808 del Parlamento Europeo y del Consejo. (14 de noviembre de 2018). Por la que se modifica la Directiva 2010/13/UE sobre la coordinación de determinadas disposiciones legales, reglamentarias y administrativas de los Estados miembros relativas a la prestación de servicios de comunicación audiovisual (Directiva de servicios de comunicación audiovisual), habida cuenta de la evolución de las realidades del mercado. http://data.europa.eu/eli/dir/2018/1808/oj

Otra documentación

Bill C-63: An Act to enact the Online Harms Act, to amend the Criminal Code, the Canadian Human Rights Act and An Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service and to make consequential and related amendments to other Acts. (2021). 1st Reading, Feb. 26, 2024, 44th Parliament, 1st sesión. https://www.parl.ca/LegisInfo/en/bill/44-1/C-63

Cámara de Diputadas y Diputados. (10 de octubre de 2018). Boletín 12164-07. Modifica el Código Penal con el objeto de sancionar la difusión no consentida de material con connotación o de índole sexual.