LA LIBERTAD DE EXPRESIÓN ANTE LA SOCIEDAD ALGORÍTMICA: UNA MIRADA AL PROBLEMA DE LA MODERACIÓN DE CONTENIDOS ONLINE*

FREEDOM OF EXPRESSION IN THE FACE OF THE ALGORITHMIC SOCIETY: AN APPROACH TO ONLINE CONTENT MODERATION PROBLEM

Working Paper IE Law School AJ8-259 11-12-2020
Cátedra José María Cervelló

Juan Albero Valdés
Abogado
jalberovaldes@gmail.com

Resumen. La irrupción de las plataformas online nos ofrece un sinfín de oportunidades para promover una libertad de expresión robusta y vibrante. Sin embargo, pese a este cariz positivo y esperanzador, también comporta serios riesgos y desafíos para la misma. Precisamente, la capacidad de estas compañías tecnológicas para afectar y modelar el discurso público mediante la moderación de contenidos online pone en grave peligro nuestra cultura democrática. Este ensayo tiene como objetivo analizar el fenómeno y problema de la moderación de contenidos, sus implicaciones para la libertad de expresión y sus posibles soluciones. Asimismo, también reflexiona respecto al rol y responsabilidades de las propias plataformas online en el marco de la Directiva de Comercio Electrónico.

Palabras clave: libertad de expresión, moderación de contenidos, regulación de contenidos, plataformas online, intermediarios, algoritmos

Abstract.*The irruption of online platforms offers us endless opportunities to promote robust and vibrant freedom of expression. However, despite this positive and hopeful aspect, it also entails serious risks and challenges. Indeed, the ability of these technology companies to affect and shape public speech through online content moderation puts our democratic culture at serious risk. This essay aims to analyze the phenomenon and problem of content moderation, its implications for freedom of expression and posible solutions. Likewise, it also reflects on the role and responsabilities of the online platforms themselves within the framework of the Ecommerce Directive.*

Keywords: *freedom of expression, content moderation, content regulation, online platforms, intermediaries, algorithms*

** Este trabajo fue ganador del XIV Premio Cervelló de Derecho de los Negocios 2020.*

1. Introducción

El advenimiento de la sociedad algorítmica ha puesto de manifiesto la dificultad de materializar el *desiderátum* de crear "un mundo en el que cualquier persona, en cualquier lugar podrá expresar sus creencias, sin importar cuán singulares sean, sin miedo a ser coaccionado al silencio o el conformismo" (Barlow, 1996, p. 241). Lejos de lograr tal cosa, esta sociedad, caracterizada por la aparición de grandes plataformas tecnológicas que se han ubicado entre los Estados y los individuos, por la utilización de la inteligencia artificial (IA) y los algoritmos en la toma de decisiones, y, cuya savia vital son los datos, presenta, más bien, serios desafíos para la libertad de expresión de los ciudadanos.

Precisamente, el 28 de mayo de 2020 el presidente de los Estados Unidos, Donald Trump aprobó la *Executive Order on Preventing Online Censorship*, documento en el cual se pusieron en evidencia dos cuestiones. Primero, la creciente preocupación que despiertan los gigantes tecnológicos por su capacidad de afectar y modelar el discurso público y, segundo, la necesidad de abrir un debate acerca de las reglas y responsabilidades que deben imponerse a las plataformas en relación con los contenidos que sus usuarios publican y difunden a través de ellas. Y es que, sin lugar a dudas, en la actualidad, los límites del debate público están siendo cada vez más condicionados por las plataformas online debido, básicamente, a una prerrogativa: el poder de moderar.

Pero, el atributo de la moderación acentúa una curiosa paradoja: mientras que por un lado las plataformas actúan como auténticos catalizadores de la libertad de expresión al crear amplios foros de debate para los ciudadanos; por otro lado, interfieren en ella al retirar aquellos contenidos que son considerados erróneamente como ilegales o inapropiados. Contradicción, sin duda alguna, que hace del problema de la moderación un asunto complejo de resolver. Este ensayo, precisamente, busca abordar esta cuestión: *¿cómo es posible articular una moderación de contenidos que salvaguarde al mismo tiempo la libertad de expresión de los usuarios de Internet?*

En aras de ofrecer una respuesta a esta pregunta, el objetivo de este trabajo consiste en analizar el problema de la moderación, sus implicaciones para la libertad de expresión y sus posibles soluciones. Para alcanzar tal propósito, este ensayo se divide en tres partes. En la primera de ellas se abordará el fenómeno y problema de la moderación. En la segunda, analizaremos la regulación de contenidos y, en especial, el régimen de responsabilidad de los intermediarios online en el marco de la Directiva de Comercio Electrónico. Finalmente, en la última parte se propondrán diversas soluciones para, de alguna manera, hacer de la moderación una actividad que resulte coherente con el respeto a la libertad de expresión de los usuarios de Internet.

2. El poder de moderar

2.1. Concepto y características de la moderación

La moderación puede definirse como aquella práctica basada en la revisión, evaluación, categorización, aprobación y eliminación de contenidos online de acuerdo a unas normas o reglas preestablecidas. Con ella, las plataformas buscan apoyar e incentivar los comportamientos comunicativos positivos en la red, a la vez que intentan minimizar las actitudes agresivas, así como eliminar las publicaciones ilegales o de índole ofensiva (Flew et al, 2019). Así, en un contexto dado por la utilización masiva de las redes sociales como espacios concebidos no sólo para el intercambio de fotos y videos, sino también como verdaderos foros de discusión política, la moderación de contenidos se erige como un auténtico instrumento de control y organización del discurso público.

Sentado lo anterior, cabe pasar a exponer a continuación cuáles son las características más relevantes de esta práctica. En primer lugar, la actividad moderadora puede realizarse antes de la publicación de contenidos en la red (moderación *ex ante*) o después (moderación *ex post*). Asimismo, podemos hablar también de moderación *reactiva* si esta se produce como consecuencia de una notificación por parte de un tercero o de moderación *proactiva* si es la plataforma la que inicia la moderación por su propia voluntad. En segundo lugar, la moderación de contenidos puede ejecutarse de tres formas distintas: i) *manual*, mediante el empleo de equipos de trabajadores entrenados para revisar y, en su caso, retirar los contenidos; ii) *automatizada,* a través de la implantación de herramientas tecnológicas para filtrar, detectar, separar y retirar los contenidos; e iii) *híbrida,* mediante la incorporación de elementos procedentes tanto de la moderación manual como de la automatizada. Generalmente, a la hora de moderar contenidos la forma hibrida es la más utilizada tanto por las plataformas grandes como por las pequeñas (Singh, 2019). Ciertamente, dadas las particularidades que presentan las plataformas – grandes volúmenes de datos, velocidad de las comunicaciones y facilidad para acceder a las mismas –, el empleo de las tecnologías resulta esencial para acometer la actividad de moderar a gran escala. De ahí que, cada vez más, plataformas como Facebook o Twitter estén basando sus procesos de moderación en sistemas algorítmicos y en el uso intensivo de la IA con el fin de obtener mejores resultados a la hora de filtrar y retirar contenidos relacionados con, por ejemplo, el discurso de odio o el terrorismo (Citron y Jurecic, 2018).

En tercer lugar, dependiendo del grado de control que se ejerza sobre los contenidos, cabe distinguir entre la moderación que determina qué contenidos son o no aceptables para ser publicados (*hard control*), de aquella que simplemente organiza y prioriza los contenidos que los usuarios visualizan (*soft control*) (York y Zuckerman, 2019). La primera de ellas se asocia con la función de *gatekeepers*, esto es, establecer qué categoría de contenidos están permitidos o no en la plataforma. Por el contrario, la segunda forma se vincula con la función de organizador, en la medida en que la plataforma busca personalizar y simplificar la navegación de los usuarios en la red mediante la organización y disposición de los contenidos de acuerdo a sus propias preferencias y gustos personales, para así, hacer de la participación de aquellos, una experiencia (Gillespie, 2018 a).

En la práctica, el gobierno de las plataformas reside en tres instrumentos; los Términos y Condiciones de Servicio ("TCS"), los cuales constituyen un contrato de adhesión que suscriben y aceptan los usuarios al registrarse; las Directrices de Contenidos ("DC"), a través de las cuales se establecen ciertos estándares sobre qué tipo de discurso es aceptable para la plataforma, y, el proceso de moderación de contenidos, mediante el cual se ejecutan los TCS y las DC (Flew et al, 2019). En un primer momento, compañías como Facebook o YouTube carecían de un *corpus normativo* que regulase la actividad de moderación por lo que los trabajadores encargados de moderar los contenidos apenas disponían de reglas internas en las que apoyarse a la hora de proceder a la retirada de las publicaciones. Estas reglas, también denominadas como "*community standarts*", constituían principios vagos, ambiguos y poco desarrollados, por lo que su aplicación derivó muchas veces en problemas de interpretación, arbitrariedad y discriminación. Posteriormente, con el aumento del número de los usuarios y del volumen de publicaciones, el sistema de *community standarts* dejó paso a un complejo sistema de reglas: los TCS y las DC (Klonick, 2018).

Así las cosas, los TCS constituyen la relación contractual entre los usuarios y las plataformas por lo que contienen, junto con las DC, las reglas que determinan qué contenidos deben o no ser moderados y de qué manera, es decir, regulan el discurso de los usuarios en la plataforma. Como consecuencia de la diversidad, dimensión y naturaleza global de la comunidad a la que regulan, los TCS y las DC no reflejan necesariamente ningún sistema legal específico, si bien es cierto que, en la medida en que se diseñan para prevenir posibles daños y abusos, estas reglas internas se superponen muchas veces a las leyes nacionales (De Streel et al, 2020). Debe señalarse, por consiguiente, que las plataformas no sólo moderan aquellos contenidos que son ilegales, sino también aquellos que, siendo legales, son considerados como ofensivos o inapropiados según los TCS y las DC.

2.2. Los fundamentos de la moderación: ¿por qué moderan las plataformas?

Desde un primer momento, el ingente volumen de contenidos que se publicaban en la red, la velocidad con que estos se intercambiaban entre los usuarios y la posibilidad de realizar todas estas acciones de una manera anónima[1], pusieron en evidencia la dificultad de concebir el ciberespacio desde una perspectiva puramente de *laissez faire*; se constató, por el contrario, la necesidad de establecer ciertos controles que permitiesen desarrollar un espacio de comunicación seguro y viable. Esta necesidad, sin embargo, debía ser congruente, al mismo tiempo, con el imperativo de incentivar la libertad de expresión y desarrollar un ecosistema digital robusto. Ambas exigencias se vieron materializadas con la aprobación por el congreso de los Estados Unidos en 1996 de la norma que contribuiría de una manera extraordinaria a la

[1] Precisamente, como señala L. Lindsky, diversos estudios revelan que los usuarios son más propensos a desarrollar actitudes soeces y abusivas cuando la comunicación esta mediada por la tecnología, y, en particular, cuando esta permite el anonimato. En este sentido, el empleo de la tecnología impone una separación entre el orador y su audiencia que termina creando un efecto "desinhibidor". Efecto que supone tanto una virtud como un vicio para la expresión online; virtud, por cuanto contribuye al desarrollo de un discurso abierto y robusto; vicio, en la medida en que favorece que este sea soez y abusivo. (Lidsky, 2011).

promoción de la cultura de la libertad de expresión y al crecimiento de Internet; la sección 230 de la *Comunication Decency Act* ("CDA") [2].

En efecto, la sección 230 de la CDA introduce una doble inmunidad para las plataformas; por un lado, una inmunidad por los contenidos publicados por terceros y, por otro lado, una inmunidad por la retirada o permanencia de dichos contenidos. Esta última, amparada en la cláusula "buen samaritano" constituye un poderoso acicate para que las plataformas lleven a cabo la moderación de contenidos. Precisamente, el legislador consideraba que sin dichas inmunidades las plataformas se verían atrapadas en lo que algunos denominaron como el "dilema del moderador": reticencia a participar en la moderación de contenidos para evitar ser tratados por los tribunales como editores con responsabilidad legal por la no retirada de materiales ilegales (Keller, 2020; Skorup y Huddleston 2019).

Concretamente, este dilema tuvo su punto de inflexión en el asunto *Stratton Oakmont, Inc. v. Prodigy Servs. Co.*[3], al responsabilizarse a la compañía Prodigy por los comentarios publicados por un tercero, habida cuenta de que la corte de apelación estimó que, al controlar los comentarios, la plataforma no actuaba como un mero distribuidor, sino como un editor. Posteriormente, en el asunto *Zeran vs. American Online, Inc*. [4], la corte de apelación señaló que la sección 230 había sido promulgada con el fin de eliminar los desincentivos para la autorregulación que había creado el caso *Stratton Oakmont*. Así, en sus orígenes la actividad de moderación se concebía, bajo el amparo de la sección 230, como un instrumento de autorregulación destinado a fomentar la utilización por parte de las plataformas de tecnologías que bloquearan y filtraran los contenidos ofensivos publicados por los usuarios (Keats y Wittes, 2017).

Sin embargo, crear un foro de debate viable, minimizando los contenidos ofensivos y en el que los usuarios puedan publicar y compartir ideas no es el único motivo que lleva a las plataformas a practicar la moderación de contenidos. Efectivamente, las plataformas buscan, además, crear ecosistemas digitales saludables que sean favorables para el desarrollo de redes y canales comerciales con el fin de promover la publicidad y, por tanto, obtener mayores ingresos (De Gregorio, 2019). El modelo de ingresos sobre el que se asientan las plataformas depende del flujo de contenidos generados por los usuarios y de las interacciones que se generen en la propia plataforma entre los demandantes, los oferentes y los anunciantes de bienes y servicios (Van Hoboken et al, 2018). De tal forma que, a mayor cantidad de contenidos generados y de interacciones, mayor perspectiva de ganancias por publicidad. Un claro ejemplo de este poder de condicionar la moderación de contenidos se puso de manifiesto en 2017, cuando varios informes sobre la existencia de contenido ofensivo en YouTube hicieron a los anunciantes amenazar a la plataforma con la retirada de sus anuncios. Ante esto, YouTube se vio obligada a cambiar sus políticas de contenidos (Keller, 2018). Sin embargo, no siempre la existencia de contenidos ofensivos o de carácter controvertido puede suponer una amenaza para este modelo de negocio. En efecto, las plataformas pueden tener un incentivo para permitir la publicación de la desinformación o de ciertos contenidos ofensivos, siempre y cuando dicho material sirva para mantener la atención y participación de los usuarios en la plataforma y, de ese modo, conseguir un mayor número de ingresos. Esto, quedó suficientemente claro durante las

[2] 47 U.S. Code § 230.
[3] Asunto *Stratton Oakmont, Inc. v. Prodigy Servs. Co*., 1995 WL 323710 (N.Y.Sup.Ct. May 24, 1995).
[4] Asunto *Zeran vs. American Online, Inc.,* 129 F.3d 327 (4th Cir. 1997).

elecciones estadounidenses de 2016, en las cuales la utilización de *fake news*, de cuentas "*bot*" automáticas y otros métodos de manipulación se realizó con la aquiescencia de algunas plataformas, a pesar de ser perjudiciales y de tener un impacto negativo para la formación de la opinión pública (Helberger et al, 2018).

Si bien es cierto que las razones económicas constituyen uno de los principales motivos para moderar, cabe señalar otras causas por las cuales las plataformas se ven abocadas a llevar a cabo tal práctica, aunque sea de una manera menos directa, a saber: la filosofía corporativa de la compañía, en la medida que responden ante sus accionistas; la presión de la opinión pública, al verse afectadas por las demandas sociales dirigidas a promover la retirada de contenidos nocivos u ofensivos; y, el marco legislativo del país donde operen (Sander, 2020).

2.3. El problema de la moderación

2.3.1. La construcción del discurso y su impacto en la libertad de expresión

La "plataforma", como metáfora que es, no sólo evoca en el imaginario del usuario una infinidad de usos, beneficios y oportunidades, sino que también encierra, bajo la promesa de un espacio abierto y libre, una heterogénea cantidad de riesgos y amenazas no tan perceptibles para el mismo. Las plataformas online no median el discurso, sino que lo constituyen (Gillespie, 2018 b). Cuando Jack Dorsey, CEO de Twitter, se refirió a la misma como "*digital public square*", es decir, como un espacio en el que intercambiar *libre* y *abiertamente* todo tipo de contenidos (Brannon, 2019), no estaba definiendo su compañía erróneamente, pero si de una forma inexacta. Tal inexactitud radica en el hecho de que las plataformas no son meros intermediarios del discurso, sino que contribuyen a crear dicho discurso a través de la moderación, sobre la base de sus reglas internas (TCS y DC) y de conformidad con las exigencias derivadas de su modelo de negocio. En efecto, qué duda cabe de que, al acceder a una plataforma, todo usuario se somete a unas condiciones de uso y a unas reglas de comportamiento que no existen, por el contrario, en un espacio público.

Sin embargo, la *construcción* del discurso público a través de la moderación de contenidos entraña, a su vez, ciertos riesgos y amenazas para los derechos fundamentales, como es el caso de la libertad de expresión. Precisamente, al analizar el impacto del ecosistema digital en el discurso, hay quienes sostienen que más que una vulneración de la libertad de expresión debería hablarse de una libertad de expresión menos *libre.* Es lo que Yemini (2018) ha denominado como "la ironía de la libertad de expresión". Según éste, la disrupción tecnológica ha conducido a la libertad de expresión en direcciones contradictorias: mientras que por un lado ofrece a los usuarios una capacidad de expresión casi ilimitada, por otro lado, disminuye su libertad de expresarse. Esto, es lo que llamamos como problema de la moderación. La posible afectación de la libertad de expresión como consecuencia de la implementación de la moderación de contenidos. En otras palabras, ¿cómo configurar una serie de procesos de control y organización de contenidos que respeten la libertad de expresión? Esta circunstancia, a su vez, sitúa a las plataformas en el intrincado dilema de articular espacios seguros y atractivos para los usuarios, pero sin que al mismo tiempo esto suponga un menoscabo de su libertad de expresión. Un complejo equilibrio de intereses que se ha visto condicionado por aquellos que afirman que las

plataformas moderan demasiado y, aquellos otros que, en cambio, abogan por una moderación más agresiva contra la proliferación de contenidos ilícitos. Probablemente, las plataformas moderen demasiado poco algunas veces y bastante, otras. Según el informe *Free Espression, Harmful Speech and Censorship in a Digital World*, la mayoría de estadounidenses considera que las plataformas no hacen suficiente para eliminar los contenidos ofensivos y solamente uno de cada cinco valora la moderación de las plataformas como demasiado severa (Knight Foundation y Gallup, 2020).

Pues bien, tales riesgos y amenazas para la libertad de expresión son consecuencia de las limitaciones que presenta la moderación de contenidos, y, en particular, de la utilización de medios automatizados. En este sentido, resulta apropiado traer a colación el *Libro Blanco sobre la Inteligencia artificial* (Comisión, 2020) y el *Informe sobre la promoción y protección de la libertad de expresión* (Relator Especial de las Naciones Unidas, 2018), los cuales advierten de los posibles daños que pueden ocasionar el empleo de IA y los algoritmos en la libertad de expresión. Cuando hablamos de *limitaciones* hacemos referencia a aquellos elementos o esferas de preocupación que por sus características y naturaleza pueden afectar a la libertad de expresión de los usuarios. A efectos de analizar esta cuestión, cabe diferenciar entre limitaciones estructurales, materiales y procedimentales.

2.3.2. Limitaciones estructurales

Una de las principales notas características de las plataformas es su capacidad de conectar grandes comunidades de usuarios, lo que supone en la práctica la exigencia de almacenar, supervisar y gestionar ingentes cantidades de datos a nivel global. Sin embargo, dicha capacidad de conectar no significa capacidad de controlar. Las plataformas no pueden, debido a su estructura y recursos limitados, monitorizar y controlar la totalidad de los contenidos publicados. Esta limitación, supone un notable inconveniente para examinar de forma pormenorizada la aceptabilidad de un contenido, circunstancia que conlleva, por consiguiente, la tendencia a aplicar análisis generalistas.

Por otro lado, la exigencia de moderar a gran escala significa realizar la compleja tarea de organizar y establecer las reglas del discurso en comunidades conformadas por usuarios de distintas nacionalidades, religiones, culturas e ideologías. Esto, supone para las plataformas que actúan en el plano internacional, hacer frente a presiones globales a la hora de moderar. Presiones que, dadas las limitaciones estructurares que presentan las plataformas, resultan cuanto menos difíciles de gestionar. En concreto, esta cuestión puede situar a las plataformas ante la difícil situación de elegir entre buscar una moderación por diseño, en la que se realicen análisis más particulares y se tengan en cuenta los aspectos socio-culturales de las regiones donde se opere, pero mucho más costosa y difícil de aplicar, o, por el contrario, verse abocadas a implementar una aproximación generalista, menos costosa pero más susceptible de ocasionar restricciones en la libertad de expresión.

2.3.3. Limitaciones materiales

En términos de limitaciones materiales, las plataformas han sido criticadas por aplicar reglas internas que resultan ser en la práctica, bien excesivamente restrictivas, bien demasiado permisivas (Sander, 2020). Los motivos que subyacen a esta crítica tienen que ver con la complejidad que reviste muchas veces la ejecución y aplicación de las reglas internas al caso concreto. En este sentido, un primer aspecto a destacar está relacionado con la interpretación del contenido cuya licitud se discute. En efecto, los algoritmos no pueden interpretar el tono en que se emite el mensaje ni contextualizar el contenido que se publica. Tampoco pueden identificar emociones, sarcasmos o ironías. Ni distinguir una imagen de desnudos con valor histórico, cultural o didáctico, de una mera fotografía de carácter obsceno. Precisamente, en septiembre de 2016, Facebook retiró un post del periodista Tom Egeland en el que se publicaba la fotografía *The Terror of the War*, más conocida comúnmente como la "Niña de Napalm". Esta imagen, galardona en 1972 con el Premio Pulitzer, constituye una de las representaciones más icónicas y críticas con la Guerra de Vietnam; sin embargo, para la red social aquella no cumplía con sus normas de contenidos (Gillespie, 2018 b). Un caso similar ocurrió con el periódico estadounidense *The Vindicator*, cuya publicación de un fragmento de la Declaración de Independencia en la víspera del 4 de julio de 2018 fue retirada por Facebook, al considerar que la misma violaba sus reglas sobre discurso de odio (Aswad, 2018).

Sin embargo, tal déficit interpretativo y de contextualización no resulta únicamente atribuible a los sistemas algorítmicos, sino que también puede hallarse en los moderadores humanos. Por ejemplo, Nicole Wong y su equipo de moderadores de YouTube retiraron en el año 2007 un vídeo en el que aparecía un hombre siendo golpeado brutalmente en una celda. Sin embargo, el video no había sido publicado con el fin de difundir contenidos violentos, sino por un activista con el propósito de denunciar ante la comunidad internacional la violación de los derechos humanos por parte de la policía egipcia. Ante las críticas recibidas, YouTube restauró el video (Klonick, 2018).

Por otro lado, las reglas internas de las plataformas que prohíben la publicación de determinados contenidos, resultan ser, muchas veces, vagas, ambiguas e imprecisas. En particular, esta problemática surge a la hora de definir conceptos sobre los cuales no hay una definición clara y concreta al respecto, como puede ocurrir con palabras como "extremista", "terrorismo" o "violencia", máxime cuando las plataformas operan en diversos países y deben someterse a ordenamientos jurídicos y sistemas de valores muy distintos entre sí. Esta vaguedad a la hora de definir la ilegalidad de los contenidos desplaza sobre las plataformas la carga de dilucidar la intención o *telos* del emisor del contenido (De Streel et al, 2020).

Sin embargo, aquello que más dificultades presenta es determinar si la publicación de un contenido resulta aceptable o no. No obstante, tal valoración puede entrañar una mayor o menor complejidad dependiendo de la tipología del contenido. Así, en cuestiones de abuso sexual a menores, pornografía o incluso copyright, la retirada del contenido, en caso de considerarse el mismo como ilícito, no presenta excesiva dificultad. Sin embargo, hay otras veces que, dicha decisión no resulta sencilla, sino que requiere, por el contrario, un análisis riguroso y pormenorizado del asunto en cuestión; análisis que excede de las capacidades no sólo de un algoritmo, sino también de un moderador humano. Hacemos alusión a cuestiones relacionadas con el discurso del odio, la incitación a la violencia o el terrorismo. Por ejemplo, en el año 2018,

Facebook consintió la publicación de un comentario en el que se afirmaba la negación del Holocausto por parte de un usuario. La decisión adoptada por la plataforma desencadenó multitud de críticas hacia la misma al considerarse que la publicación constituía claramente un mensaje de odio (Aswald, 2018). Otro de los problemas que pueden darse a la hora de valorar contenidos es la discriminación de colectivos y grupos minoritarios, etnias o comunidades lingüísticas con menor presencia. Ello es debido, generalmente, a la utilización de tecnologías como, por ejemplo, el *machine learning*. Si los datos están influenciados por sesgos o desigualdades del mundo real, en el momento de analizarse estos, la tecnología puede llevar a reflejar o amplificar dichas desigualdades, lo que puede ocasionar un silenciamiento ilegítimo para aquellos (Llansó et al, 2020).

Como resultado de estas limitaciones, la moderación de contenidos y, en particular, la utilización de medios automatizados, adolecen de falsos positivos (algo que es erróneamente clasificado como inaceptable) y negativos (no se clasifica como inaceptable algo que debería haber sido clasificado como tal) a la hora de detectar los contenidos ilegales e inapropiados. Precisamente, son los falsos positivos los que mayores repercusiones presentan para la libertad de expresión puesto que pueden actuar como un elemento de censura al eliminar contenidos que, considerándose por el algoritmo como ilícitos o inapropiados, no los son en realidad (Llansó et al, 2020). Por tanto, se atribuye a los algoritmos una responsabilidad excesiva a la hora de determinar qué contenidos pueden o no ser retirados de la plataforma. Esto supone que los resultados que se obtengan en materia de eliminación de contenidos variarán en función de la tecnología que se disponga, lo que en la práctica implica que aquellas compañías que no puedan invertir en mejoras de tecnología, probablemente se verán obligadas a tolerar altas tasas de falsos positivos para evitar la responsabilidad por los falsos negativos (Keller, 2018).

Pero, el problema central que presentan los algoritmos ya no es que ofrezcan un porcentaje mayor o menor de aciertos a la hora de retirar los contenidos potencialmente ofensivos, sino que no pueden substituir el juicio humano en cuestiones concernientes a valoraciones de carácter jurídico. En este sentido, las plataformas han conferido la salvaguarda de los derechos y libertades de los usuarios a los algoritmos, situación que ha conducido a una "matematización de la ley", desde el mismo momento en que son aquellos los que contribuyen a definir el concepto de legalidad (De Gregorio, 2019).

2.3.4. Limitaciones procedimentales

Desde un punto de vista procedimental, la moderación es el conjunto de decisiones que se adoptan para controlar y organizar los contenidos en la plataforma. En este sentido, el principal problema que presenta la moderación es la opacidad a la hora de realizar tales procedimientos. Esta limitación es consecuencia, en gran parte, de la relación asimétrica que une al usuario con la plataforma (De Gregorio, 2019; Balkin, 2018). En efecto, mientras que la plataforma almacena, analiza y retira los contenidos de los usuarios, estos últimos apenas conocen la manera en que aquellos son gestionados o los motivos por los cuales acaban siendo bloqueados. Esto, no sólo conduce a confusión entre los usuarios, sino que también dificulta tener un debate público informado sobre cómo regular el contenido de Internet de una manera que se salvaguarde la libertad de expresión (Suzor et al, 2019).

Es cierto, no obstante, que desde que Google publicó su primer informe de transparencia en 2010, el número de compañías que se han adherido a esta buena práctica ha aumentado considerablemente cada año (Sander, 2020). Y que, además, con la publicación de los Principios de Santa Clara (2018) se ha dado un paso más a la hora de exigir a las plataformas que faciliten más información sobre cómo moderan. En particular, mediante estos principios se pretende mejorar la transparencia en dos niveles: a nivel individual, a través de la información que debe facilitarse a los usuarios y, a nivel de la plataforma, mediante la publicación de información agregada (Suzor et al, 2019). Sin embargo, según el informe *Ranking Digital Rights Corporate Accountability Index* (2019), el nivel de transparencia sigue siendo insuficiente. Este informe resulta relevante desde un punto de vista analítico en la medida en que examina el grado de transparencia que presentan las plataformas en materia de políticas y prácticas que afectan a la libertad de expresión sobre la base de varios indicadores, entre ellos, la accesibilidad y claridad de los TCS, las restricciones de contenidos o las solicitudes por parte de los gobiernos o terceros. Así, uno de los principales ámbitos de opacidad que afectan a las plataformas es la falta de información sobre cómo se aplican los TCS, esto es, sobre el volumen y naturaleza de las acciones adoptadas por las plataformas para restringir los contenidos.

A lo anterior, cabe añadir la total opacidad que presentan los medios automatizados, en particular, los sistemas algoritmos. Precisamente, el desconocimiento de su codificación, de las bases de datos que emplean o de cómo funcionan sus procesos de toma de decisiones convierten a estos sistemas en auténticas "*black boxes*" (Sigh, 2019; De Gregorio, 2019). Hay quienes sostienen que dicha falta de transparencia no es resultado de la complejidad de la moderación ni tampoco de una circunstancia casual, sino, más bien, de una actuación deliberada. Según Roberts (2019), esta disposición a omitir y ocultar información es resultado de lo que llama "lógica de la opacidad" cuya finalidad no es otra que presentar las plataformas como elementos objetivos y neutrales en el imaginario del público. Esta lógica es un acto de "despolitización" en la medida en que el contenido publicado por los usuarios tiene para las plataformas un único valor que varía dependiendo de su capacidad de atraer usuarios y relacionarlos con anunciantes. En definitiva, para Roberts, la falta de transparencia sirve a fines estrictamente económicos.

3. La regulación de contenidos: de los puertos seguros a las medidas proactivas

3.1. El modelo de regulación comunitario y las iniciativas públicas

3.1.1. El régimen de responsabilidad de los intermediarios a la luz de la Directiva de Comercio Electrónico

Una vez planteado y delimitado el problema de la moderación, conviene abordar cual es el rol y las responsabilidades de las plataformas o intermediarios, por los contenidos ilegales que almacenan y circulan por sus sitios web. Siguiendo el enfoque regulatorio que habían iniciado los Estados Unidos en 1996 con la aprobación de la sección 230 de la CDA y, posteriormente,

con la promulgación de la *Digital Millenium Copyright Act* ("DMCA")[5], la Unión Europea abordó la cuestión de la responsabilidad de los intermediarios con la Directiva 2000/31, de 8 de junio de 2000, sobre comercio electrónico ("DCE")[6], la cual constituye un régimen horizontal aplicable a cualquier tipo de contenido ilegal.

De forma similar a la DMCA, la DCE establece en sus artículos 12 a 14 un sistema de exención de responsabilidad condicionada (*safe harbours)* para tres grupos de intermediarios dependiendo de la tipología de servicios que presten: mera trasmisión, memoria tampón (*caching)* y almacenamiento (*hosting)*. Asimismo, el artículo 15 de la DCE establece la prohibición de imponer una obligación general de supervisión a los intermediarios. Como bien ha tenido ocasión de señalar el Tribunal de Justicia de la Unión Europea ("TJUE"), la DCE no establece los requisitos para que se declare la existencia de responsabilidad, pues estos se encuentran en la legislación nacional, sino que sirve, por el contrario, para restringir los supuestos en los que, conforme a la legislación nacional aplicable, puede generarse responsabilidad para los intermediarios[7]. Puesto que la gran mayoría de las actividades realizadas por los intermediarios se encuadran dentro de los servicios de almacenamiento o *hosting* (Van Hoboken et al, 2018) y que, como ha señalado el TJUE en el asunto C-360/10 (*Netlog*)[8], las plataformas de redes sociales en línea constituyen servicios de *hosting*, centraremos nuestro estudio en este tipo de intermediarios.

3.1.2. Las iniciativas públicas contra la difusión de contenidos ilegales

Previo al análisis de la responsabilidad de los intermediarios de servicios de *hosting*, conviene exponer las diferentes iniciativas de carácter legislativo y no legislativo que se han llevado a cabo en los últimos años, tanto por parte de la Comisión Europea como por parte de algunos Estados, con el fin de abordar la problemática que presenta la difusión de contenidos ilegales en las plataformas online. Estas iniciativas, que vienen a complementar el régimen previsto en la DCE, tienen por objetivo, entre otros, incentivar a estas compañías tecnológicas a que sean más transparentes y a que adopten medidas proactivas con ánimo de detectar y retirar los contenidos ilegales, principalmente, mediante la utilización de tecnologías de detección automática y filtrado.

En este contexto, la Comisión presentó en el año 2016, junto con Facebook, Microsoft, Twitter y YouTube, el Código de Conducta para la lucha contra la incitación ilegal al odio en Internet (Comisión, 2016), un documento por medio del cual dichas empresas se comprometen a adoptar determinadas actuaciones como contar con procedimientos de notificaciones claros y eficaces o revisar la mayoría de las notificaciones en menos de 24 horas. Un año después, se presentó, por un lado, la comunicación sobre la lucha contra el contenido ilícito en línea (Comisión, 2017

[5] 17 U.S. Code § 512.

[6] Directiva 2000/31/CE del Parlamento Europeo y del Consejo de 8 de junio de 2000 relativa a determinados aspectos jurídicos de los servicios de la sociedad de la información, en particular el comercio electrónico en el mercado interior (Directiva sobre el comercio electrónico), *DOCE* L 178/1, 17.7.2000.

[7] STJUE de 23 de marzo de 2010, *Google France v. Louis Vuitton et al,* C-236/08 a C-238/08, ECLI:EU:C:2010:159, ap. 107.

[8] STJUE de 16 de febrero de 2012, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) y Netlog NV,* C-360/10, ECLI:EU:C:2012:85, ap. 27.

a), cuyo contenido establecía la necesidad de que las plataformas intensificasen su lucha contra los materiales ilícitos a través de la adopción de medias proactivas eficaces, y, por otro, la *Database of Hashes*, una base de datos desarrollada en colaboración con las compañías que habían suscrito el citado código de conducta, para identificar contenidos potencialmente terroristas en las redes sociales e impedir su reaparición en otras plataformas (Comisión, 2017 b). Estas iniciativas fueron seguidas por la Recomendación sobre medidas para combatir eficazmente los contenidos ilícitos en línea (Comisión, 2018 a), documento a través del cual se invitaba a las plataformas a tomar medidas activas, proporcionadas y específicas en relación con los contenidos ilícitos. En esta línea se muestra también la Propuesta de Reglamento para la prevención de la difusión de contenidos terroristas (Comisión, 2018 b), al mencionar la idoneidad de adoptar medidas proactivas y hacer especial hincapié en la utilización de medidas automatizadas para identificar y retirar los contenidos de índole terrorista.

Junto a las iniciativas planteadas por la Comisión Europea, en el plano nacional algunos Estados también han presentado sus propias medidas para abordar la cuestión de los contenidos ilícitos. En el año 2017, Alemania aprobó la *NetzDG* o Ley de Vigilancia de la Red[9] con el fin de abordar el llamado "discurso del odio" en las plataformas online. Esta ley plantea, entre otras medidas, la obligación para aquellas plataformas de más de dos millones de usuarios en Alemania, de retirar el contenido ilegal en un plazo de 24 horas, o de 7 días para los casos que revistan mayor complejidad, bajo el riesgo de ser apercibidas con multas de hasta 50 millones de euros. De forma similar, países como Francia[10] o Reino Unido (GOV UK, 2019), también han presentado distintas iniciativas en relación con la publicación y difusión de contenidos ilícitos en las plataformas.

3.2. La responsabilidad de los intermediarios de servicios de *hosting*

El artículo 14, apartado 1 de la DCE permite que los intermediarios de servicios de *hosting* puedan acogerse al puerto seguro siempre y cuando:

> "a) el prestador de servicios no tenga conocimiento efectivo de que la actividad a la información es ilícita y, en lo que se refiere a una acción por daños y perjuicios, no tenga conocimiento de hechos o circunstancias por los que la actividad o la información revele su carácter ilícito, o de que,
>
> b) en cuanto tenga conocimiento de estos puntos, el prestador de servicios actúe con prontitud para retirar los datos o hacer que el acceso a ellos sea imposible".

Con ánimo de analizar el contenido de este precepto, procedemos a desgranar los principales aspectos y condicionantes cuya observancia resulta determinante para que las plataformas puedan acogerse a la exención. Asimismo, y sin perjuicio de lo anterior, también examinaremos el artículo 15 de la DCE.

[9] Ley para Mejorar la Aplicación de la Ley en las Redes Sociales (*Netzdg)* de 2017.
[10] *LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet* ("Ley Avia").

3.2.1. La neutralidad de los intermediarios ante la moderación de contenidos

Uno de los requisitos previstos en la DCE para que los intermediarios sean beneficiarios de la exención es la exigencia de asumir un rol "pasivo" o "neutral". La distinción entre prestadores "pasivos" o "activos" se encuentra recogida en el considerando 42 de la DCE. Aunque esta distinción hace referencia a los intermediarios de acceso y de memoria tampón y no a los de *hosting*, el TJUE en el asunto *Google France*, ha interpretado dicho considerando en el sentido de extender el alcance del mismo también a estos últimos. De este modo, el Tribunal señala que, para comprobar la responsabilidad del intermediario, "es necesario examinar si el papel desempeñado por el prestador es neutro, es decir, si su comportamiento es meramente técnico, automático y pasivo, lo que implica que no tiene conocimiento ni control de la información que almacena"[11].

Desde la perspectiva de la moderación de contenidos, lo anterior supone ciertos problemas, máxime cuando la propia actividad de moderar supone para las plataformas experimentar un rol activo. En este sentido, cabe plantearnos si la moderación de contenidos puede ser un obstáculo para que los intermediarios sean considerados como prestadores de servicios "pasivos" o "neutros" en el sentido del considerando 42 de la DCE. Para abordar esta cuestión resulta adecuado examinar el concepto de "control", cuya interpretación presenta *prima facie* aspectos dificultosos en términos de graduación y especificación. En efecto, el considerando no aclara - ni tampoco el TJUE- qué nivel de control debe ejercerse para ser considerado como prestador "activo". Teniendo en cuenta que los intermediarios de *hosting* desarrollan, *de facto*, un cierto nivel de control sobre la información que almacenan, cabe suponer que, el nivel de control al que se hace referencia debe ser superior al inherente en un servicio de *hosting*. Ahora bien, dónde situar dicho grado de control inherente puede resultar una cuestión espinosa dada la variada tipología y naturaleza de las plataformas que coexisten en el mercado actual.

Por otra parte, la falta de especificación en cuanto al tipo de control al que hace referencia el considerando genera dudas interpretativas. Así pues, el concepto de "control" podría asimilarse al "control" que se prevé en el artículo 14.2 de la DCE, cuyo contenido establece que "el apartado 1 no se aplicará cuando el destinatario del servicio actúe bajo la autoridad o *control* del prestador de servicios" (énfasis añadido). En cambio, otra opción podría ser interpretar el concepto de "control" como "control editorial". Ello nos conduce a plantearnos si la moderación puede ser equiparada a esto último. Si bien es cierto que la práctica de la moderación supone controlar y organizar los contenidos, no podemos afirmar que tal comportamiento se asimile al de un editor. La razón reside en el hecho de que, a través de la moderación, las plataformas no hacen suyo el contenido que almacenan, sino que simplemente lo condicionan. Incluso en el caso de que se afirmase que el contenido generado por los usuarios de conformidad con unos TCS ha sido creado bajo el control del prestador, no estaríamos ante un supuesto de control editorial (Arroyo, 2020).

Sin embargo, a pesar de la falta de semejanza entre las actividades de moderación y de control editorial, cabe el riesgo de que *lege lata* la primera pueda ser asimilada a la segunda (Arroyo, 2020). Esto no solo produce un cierto grado de incertidumbre en las plataformas, sino que, además, implica resolver la cuestión atendiendo a las circunstancias particulares del caso y a la

[11] Asunto *Google France v. Louis Vuitton et al.,* C-236/08 a C-238/08, ap. 114.

tipología concreta de servicios que preste el intermediario. Por ejemplo, en el asunto *L'Oreal* el TJUE estableció que "el mero hecho de que el operador de un mercado electrónico almacene en su servidor ofertas de venta, determine las condiciones de su servicio, sea remunerado por el mismo y dé información general a sus clientes no puede implicar que se le excluya de las exenciones de responsabilidad"[12]. En cambio, si el mismo "presta una asistencia consistente, entre otras cosas, en optimizar la presentación de las ofertas de venta en cuestión o en promover tales ofertas", debe considerarse que ha adoptado una posición activa[13].

Otra cuestión que suscita controversia tiene que ver con las iniciativas públicas para que las plataformas adopten medidas proactivas y la posibilidad de que aquellas pierdan el puerto seguro, como consecuencia de la adopción de estas. A pesar de que la Comisión (2017 a) aclara que la adopción de tales medidas no supone para las mismas desempeñar un papel activo, no hay duda de que, *de facto*, podría poner en riesgo el beneficio de la exención (Riis y Schwemer, 2019; Kaye et al, 2018).

Así pues, que una plataforma modere no implica necesariamente que, a la luz de la DCE, la misma desempeñe un rol activo que le impida acogerse a la exención prevista en el artículo 14.1 de la DCE; ahora bien, habrá que analizar el caso concreto y, en particular, la tipología de servicios que ofrezcan para determinar si la moderación que ejercen puede o no ser asimilada a un "control" sobre la información que almacenan. Ante la vaguedad de la norma y la falta de claridad por parte del TJUE, serán los tribunales nacionales quienes deberán resolver dicha cuestión. No obstante, la distinción entre intermediarios "pasivos" y "activos", además de producir cierta incertidumbre e inseguridad, no se ajusta a la realidad actual ni resulta tampoco coherente con la nueva tipología de plataformas que están surgiendo, las cuales, a través de la moderación, desarrollan conductas más activas con el fin de atraer usuarios y mejorar su experiencia.

3.2.2. El conocimiento de la ilicitud de contenidos

Una vez tratada la cuestión de la neutralidad cabe pasar a analizar las condiciones que establece el artículo 14.1 de la DCE. Los intermediarios de servicios de *hosting* no podrán ser considerados responsables por los datos almacenados siempre y cuando no tengan conocimiento efectivo de la actividad ilícita, o de que, en caso de tenerlo, actúen de manera expeditiva para proceder a su retirada. La DCE no especifica si dicho conocimiento hace referencia a un conocimiento "general" o "específico" sobre la ilicitud de los contenidos. No obstante, podemos señalar que, por "general" se alude a un conocimiento *in abstracto* de la utilización del servicio para almacenar contenido ilícito, mientras que por "específico", se hace referencia a un conocimiento *in concreto* de la existencia de un contenido ilegal (Van Hoboken et al, 2018).

[12] STJUE de 12 de julio de 2011, *L'Oréal SA, Lancôme parfums et beauté & Cie SNC, Laboratoire Garnier & Cie, L'Oréal (UK) Ltd, y eBay International AG, eBay Europe SARL, eBay (UK) Ltd, Stephen Potts, Tracy Ratchford, Marie Ormsby, James Clarke, Joanna Clarke, Glen Fox, Rukhsana Bi*, C-324/09, ECLI:EU:C:2011:474, ap. 115.
[13] Ibid., ap. 116.

La jurisprudencia comunitaria ha interpretado el concepto de "conocimiento efectivo" en el sentido de conocimiento "específico". Así, en el asunto *L'Oreal* el TJUE aclaró que un intermediario no podrá acogerse a la exención "cuando haya tenido conocimiento de hechos o circunstancias a partir de los cuales un operador económico diligente hubiera debido constatar el carácter ilícito de las ofertas de venta en cuestión"[14]. A efectos de tener conocimiento, el TJUE admitió como válida "cualquier situación en la que el prestatario en cuestión haya adquirido conocimiento, de una forma o de otra, de tales hechos o circunstancias"; esto es, como resultado de una investigación propia, de una notificación de un tercero o de la propia evidencia y notoriedad del contenido ilegal[15].

a) La diligencia del operador económico

La DCE no regula el concepto de diligencia debida, sino que lo remite en su considerando 48 al Derecho nacional, de tal forma que sean los Estados miembros quienes exijan a los intermediarios de *hosting* que apliquen dicho deber a fin de detectar y prevenir determinados tipos de actividades ilegales. Así pues, un intermediario infringe el deber de diligencia siempre que conociera o debiera conocer el contenido ilícito y, sin embargo, no hubiera actuado con el fin de retirarlo (Arroyo, 2020). No obstante, la remisión al concepto de "diligencia del operador económico" entraña varios problemas.

En primer lugar, el concepto de diligencia debida variará en función de la legislación nacional aplicable al respecto, por lo que es difícil alcanzar una definición concreta y unívoca del mismo. En segundo lugar, actuar con la diligencia debida puede exigir tanto la adopción de medidas reactivas como proactivas; a hora bien, obligar a tomar estas últimas puede tener un difícil encaje con el artículo 15 de la DCE por cuanto éste prohíbe el control general de los contenidos. No obstante, cabe señalar que la intención del legislador era fijar un deber de cuidado o *duty of care* más estrecho, con el fin de ayudar a concretar los conceptos de retirada e inhabilitación al acceso de la información ilegal, principalmente mediante la utilización de medidas reactivas (Van Hoboken, 2018). Así pues, dicho deber de diligencia podría consistir, por ejemplo, en la obligación de adoptar medidas tales como sistemas de queja o procedimientos concretos de retirada de contenidos[16]; pero, nunca, en una obligación general de supervisar los contenidos ni en un deber de vigilancia permanente.

b) La obtención del conocimiento

Básicamente, son dos los métodos a través de los cuales los intermediarios de *hosting* pueden obtener conocimiento de la ilicitud de los contenidos. El primero de ellos es *proactivo,* esto es, como resultado de las investigaciones realizadas voluntariamente por la propia plataforma. Si bien, desde la perspectiva de la responsabilidad, hay pocos incentivos para que los

[14] Ibid., ap.124.
[15] Ibid., ap. 121 y 122.
[16] Vid. Carta del Director-General del Mercado Interior, John F. Mogg, a Charlotte Cederschiöld. Bruselas, D (2000) 274, 13.06.00.

intermediarios implementen medidas proactivas para obtener el conocimiento del ilícito. Esto es debido, principalmente, al hecho de que la utilización de ciertas medidas proactivas puede poner en riesgo la neutralidad de la plataforma y, por consiguiente, suponer la pérdida de la exención. En este sentido, la falta de una cláusula de "buen samaritano" en la DCE, similar a la prevista en la CDA, contribuye a desincentivar la adopción de medidas proactivas por parte de las plataformas ante el riesgo de asumir una posición activa.

El segundo de los métodos es el *reactivo*, esto es, como consecuencia de la información proporcionada por un tercero. En particular, este método consiste en la notificación a la plataforma por parte de un tercero – un sujeto o entidad privada; una autoridad pública mediante resolución judicial o decisión administrativa; o un organismo especializado - de la existencia de contenidos ilegales. Este procedimiento recibe el nombre de *notice-and-take down* (NTD). Cabe señalarse que el individuo que realice la notificación no tiene por qué ser la víctima de la ilicitud, sino que puede ser un tercero a quien le interese la retirada de dicho contenido (Arroyo, 2020). Asimismo, algunos intermediarios han otorgado a determinadas organizaciones la condición especial de *trusted flaggers* o notificantes fiables para que les notifiquen la existencia de materiales ilícitos (ICF et al, 2018). Una vez realizada la notificación, corresponde a la plataforma examinarla. La plataforma puede, bien mantener el contenido en caso de no apreciar ilicitud alguna o bien proceder a su retirada. No obstante, también caben otras posibilidades como contactar con el proveedor del contenido para aclarar si la notificación está bien fundada o no; o para que proceda a su retirada de forma voluntaria. Con el nombre de *notice-and-action* (NA) se hace referencia a todos los procedimientos adoptados por la plataforma (incluido el NTD) con el fin de retirar los contenidos ilegales una vez recibida la notificación (Kuczerawy, 2018).

La DCE, a diferencia de la DCMA, no regula el procedimiento de NTD, sino que se limita a dejar en manos de los Estados miembros la regulación del mismo, por lo que no existe a nivel comunitario un enfoque armonizado y uniforme para la retirada de los contenidos ilegales. Este hecho ha dado lugar a que, mientras algunos Estados miembros han desarrollado un procedimiento más detallado, con un mecanismo de NTD formalizado; otros, por el contrario, se han limitado a transponer simplemente la DCE (Kuczerawy, 2017). Esta falta de regulación por parte de la DCE ha supuesto a nivel comunitario la coexistencia de procedimientos muy heterogéneos dependiendo de la tipología de contenido que se aborde (i.e. copyright, Propiedad Intelectual, pornografía infantil, etc.), de los requisitos que contenga o del grado de regulación que establezca, lo que puede suponer en la práctica un incremento de la retirada de contenidos y, por ende, una posible afectación para la libertad de expresión de los usuarios (ICF et al, 2018)[17].

[17] Precisamente, un estudio realizado en 2003 sobre el impacto que tiene el mecanismo NTD bajo la DMCA y la DCE mostraba cómo la existencia de un marco procedimental detallado, como el previsto en la DMCA, contribuía a la menor retirada de contenidos que bajo un escenario de no regulación, como es el caso de la DCE. Este estudio tuvo por objeto la creación de una página web para publicar en ella un fragmento del libro *Sobre la Libertad*, de John Stuart Mill. Una vez creada la web, los autores del proyecto enviaron una carta a una plataforma ubicada en los Estados Unidos y a otra en el Reino Unido, solicitando que el fragmento publicado fuese retirado por vulnerarse los derechos de autor. Los resultados que obtuvieron fueron distintos; mientras que en Estados Unidos la plataforma, siguiendo el procedimiento marcado por la DMCA, no retiró el contenido, la ubicada en Reino Unido lo hizo de forma automática. Debe señalarse que, *Sobre la Libertad* fue publicado en 1859, por lo que pertenece al dominio público y que los autores del proyecto enviaron las notificaciones bajo una identidad falsa, la "*John*

A la luz de lo expuesto, podemos señalar que la configuración del artículo 14.1 de la DCE presenta aspectos críticos que pueden derivar en restricciones de la libertad de expresión. En primer lugar, tener un conocimiento "efectivo" sobre una actividad ilícita supone desplazar a los intermediarios la carga de examinar el carácter de un contenido, esto es, analizar su legalidad, lo cual puede ser problemático para aquellos casos en los cuales la supuesta ilegalidad no sea tan manifiesta, máxime cuando la obtención de dicho conocimiento es, generalmente, resultado de la solicitud planteada por un sujeto privado. En segundo lugar, al establecerse la retirada expeditiva de contenidos como condición para poder acogerse a la exención, se crea un incentivo para que los intermediarios retiren sistemáticamente el material, sin realizar ninguna investigación sobre el carácter del mismo y sin permitir que el usuario cuyo contenido ha sido retirado pueda plantear las observaciones que estime oportunas al respecto. Naturalmente, esta propensión a la retirada de contenidos a la que favorece el artículo 14 de la DCE conlleva, a su vez, la eliminación de contenidos legales y legítimos, lo que puede derivar en una especie de censura privada (Kuczerawy, 2015).

3.2.3. La inexistencia de una obligación general de supervisión

El artículo 15.1 de la DCE prohíbe cualquier obligación general de supervisar los datos que los prestadores de servicios de Internet transmitan o almacenen, así como cualquier obligación general de realizar búsquedas activas de hechos o circunstancias que indiquen actividades ilícitas. Como señala el considerando 47, esta prohibición se refiere exclusivamente a obligaciones de carácter general, pero no a las específicas. En este sentido, ya en el caso *L'Oreal*, el TJUE señaló que las medidas exigidas a un intermediario no "pueden consistir en una supervisión activa del conjunto de datos de cada uno de sus clientes dirigida a evitar cualquier futura lesión de derechos"[18]. El TJUE reiteró esta interpretación en el asunto C-494/15, al aclarar que las órdenes judiciales debían ser equitativas, proporcionales y no resultar excesivamente gravosas, por lo que no puede exigirse al intermediario que realice un seguimiento general y permanente de sus clientes; ahora bien, como señala el TJUE, esto no es óbice para que se obligue al intermediario a que adopte medidas que contribuyan a evitar que se produzcan nuevas infracciones[19]. De forma análoga, dicha exégesis se confirmó en los asuntos *Scarlet* y *Netlog*, al declarar el TJUE que el requerimiento judicial por el que se ordenaba al intermediario a establecer un sistema de filtrado le obligaría a una supervisión activa de la casi totalidad de datos[20].

Sin embargo, esta línea jurisprudencial se vio alterada con la STJUE de 3 de noviembre de 2019, asunto C-18/18 (*Glawischnig*). En un caso de contenidos ofensivos y difamatorios

Stuart Mill Heritage Foundation", la cual no existe. Por tanto, solamente en que los intermediarios hubiesen realizado un breve ejercicio de investigación, habrían advertido fácilmente la imposibilidad de que la notificación prosperase (Ahlet et al, 2004). A pesar de ser un estudio a pequeña escala, sirve para poner en evidencia la importancia de contar con un marco normativo que regule de forma clara y precisa el procedimiento de NTD.

[18] Asunto *L'Oreal*, C-324/09, ap. 139.

[19] STJUE de 7 de julio de 2016, *Tommy Hilfiger Licensing LLC, Urban Trends Trading BV, Rado Uhren AG, Facton Kft., Lacoste SA, Burberry Ltd y Delta Center a.s.,* C-494/15, ECLI:EU:C:2016:528, ap. 34.

[20] STJUE de 24 de noviembre de 2011, *Scarlet Extended SA y Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM),* C-70/10, ECLI:EU:C:2011:771, ap. 40 y asunto *Netlog,* C-360/10, ap. 38.

publicados en Facebook, el TJUE interpretó el artículo 15.1 de la DCE en el sentido de que no se opone a que se pueda obligar a un intermediario a suprimir datos idénticos o similares a los declarados ilícitos con anterioridad, así como tampoco a que la retirada o bloqueo de los mismos pueda realizarse a nivel mundial[21]. Asimismo, el Tribunal admite la utilización de "técnicas e instrumentos de búsqueda automatizados" para llevar a término dichas obligaciones sin la necesidad de supervisión humana[22], lo que supone en la práctica una cuestión compleja y no exenta de crítica (Chelioudakis, 2019; Arroyo, 2020). Tal y como se expuso *supra*, los sistemas de filtrado automático no son capaces de comprender el contexto y el sentido en que se publica un contenido, aunque sea este idéntico o similar al declarado previamente como ilegal. Naturalmente, la utilización de filtrados automáticos puede dar lugar a falsos positivos y, por tanto, conllevar restricciones para la libertad de expresión de los usuarios, con riesgos particulares para aquellos ámbitos en los que se ejerza la crítica política, la denuncia social o la sátira.

3.3. *Privatised enforcement* y los riesgos de la censura privada

En el estudio *The Slide from "Self-Regulation" to Corporate Censorship,* Joe McNamme (2011) sostiene que, a la hora de analizar el modelo de autorregulación de los intermediarios online, sería más apropiado referirnos al mismo como "*devolved law enforcement*", en la medida en que son compañías privadas quienes asumen funciones policiales, judiciales y ejecutivas en relación con los incumplimientos de la ley o de sus propios TCS y DC. Esta situación, también denominada como "*privatised enforcement*" (Tworek y Leerssen, 2019; Coche, 2018; Angelopoulos et al, 2015) se ha visto incentivada por las iniciativas llevadas a cabo tanto por la Comisión como por algunos Estados para combatir la difusión de contenidos ilegales. En efecto, tanto el Código de Conducta como la *NetzDG* constituyen claros ejemplos de *privatised enforcement*.

Esta delegación en la aplicación de la ley supone que la retirada de los contenidos se realice básicamente sobre la base de las reglas internas de las plataformas. En particular, el Código de Conducta especifica que la valoración de las solicitudes se realizará "con arreglo a sus normas y directrices comunitarias y, en caso necesario, a las legislaciones nacionales de transposición de la Decisión marco 2008/913/JAJ" (Comisión, 2016, p.3). En cuanto a la *NetzDG*, los informes de transparencia aclaran que la mayor parte de la retirada de contenidos es resultado de la aplicación de los TCS y DC de las empresas, y no tanto de la normativa alemana (Tworek y Leerssen, 2019). En la práctica, tal circunstancia puede dar lugar a no pocas restricciones de la libertad de expresión, en la medida en que bajo las reglas internas de las plataformas pueden ser retirados no sólo contenidos ilegales sino también legales. Precisamente, el Código de Conducta, al animar a las plataformas a retirar el contenido relacionado con la incitación al odio sobre la base de sus TCS y DC, promueve que aquellas vayan más allá de lo fijado en la

[21] STJUE de 3 de noviembre de 2019, *Eva Glawischnig-Piesczek y Facebook Ireland Limited,* C-18/18, ECLI:EU:C:2019:821, ap. 53.

[22] La sentencia, al legitimar el uso de herramientas automatizadas sin ligar estas a la necesidad de supervisión humana podría resultar contradictoria con lo dispuesto en la Recomendación (UE) 2018/334, de la Comisión, cuyo apartado 9 establece que la utilización de estas herramientas debe contar con salvaguardias eficaces y apropiadas, como por ejemplo la vigilancia y verificación *a posteriori* por personas.

Decisión marco 2008/913/JAJ (Coche, 2018). Ello, como resultado de la vaguedad de los TCS y DC. Sin lugar a dudas, tal vaguedad permite a las plataformas disfrutar de un amplio poder discrecional a la hora de retirar los contenidos, lo que supone, muchas veces, una prerrogativa que va más allá de los objetivos iniciales, esto es, la eliminación de contenidos ilegales, provocando, a su vez, lo que algunos consideran como *censorship creep* (Keats, 2018). En este sentido se pronuncian Kaye et al, (2018) al advertir que los TCS y las DC generalmente imponen limitaciones que van más allá de lo que los Estados pueden hacer a la hora de cumplir sus obligaciones en el marco de los derechos humanos. Por otro lado, la exigencia de retirar los contenidos ilegales de la red en breves espacios de tiempo y bajo la amenaza de imponer duras sanciones, contribuye a incentivar la retirada de contenidos por parte de las plataformas, las cuales ante la imposibilidad de realizar un análisis caso por caso, prefieren retirar el contenido, independientemente de su carácter, a quedar expuestas a una posible sanción económica.

Con todo esto, es lógico sostener que la presión ejercida sobre las plataformas con el objetivo de retirar los contenidos ilegales supone hacer de las mismas, instrumentos para la implementación de las políticas públicas, con opacidad, escasa supervisión y severas implicaciones para la libertad de expresión. Concretamente, una de las principales críticas aducidas contra el modelo de *privatised enforcement* y de la presión pública tiene que ver con la posibilidad de que ambos favorezcan la censura colateral o *censorship by proxy* (Pia, 2019). Estas formas de censura privada se caracterizan por ser el Estado o un organismo público quien crea incentivos a las entidades privadas para que censuren a otros sujetos (Keats, 2018). En este sentido se pronunció Wenzel Michalski, director de *Human Rights Watch*, al afirmar que la *NetzDG* "*turns private companies into overzealous censors*" (Tworek y Leerssen, 2019, p.3). Todo ello, debe conducirnos a plantearnos si realmente este modelo de *privatised enforcement* basado en un desplazamiento exponencial de las responsabilidades sobre las plataformas tiene realmente eficacia alguna en la lucha contra la difusión de contenidos ilegales y si, además, es congruente con los principios que rigen el derecho a la libertad de expresión.

4. Repensar la moderación: de un medio de control y organización a un instrumento de garantías

4.1. Consideraciones previas

Tomando en consideración el análisis realizado, podemos sostener que, en primer lugar, los riesgos y amenazas de que se conculque la libertad de expresión en Internet no residen solamente en la propia actividad de moderar, esto es, en sus limitaciones, sino que también deben buscarse en el marco normativo que regula la moderación y en las diversas iniciativas públicas planteadas. En segundo lugar, el fenómeno y problema de la moderación refleja con claridad el modelo triádico del discurso que caracteriza a la sociedad algorítmica (Balkin, 2018), en la medida en que el propio discurso es fruto de la interrelación dinámica que se da entre las plataformas, los Estados y los usuarios. Esto significa que, la aproximación al problema planteado debe pasar, ante todo, por la cooperación entre todos ellos. Incluso, desde un punto de vista moral, podría sostenerse que, como consecuencia de esta relación triádica, la

solución debe plantearse sobre la base de una responsabilidad compartida o "corresponsabilidad" (Helberger et al, 2018).

Asimismo, debemos advertir que, dado el ingente volumen de contenidos que circulan por las plataformas, un cierto grado de error a la hora de moderar resulta inevitable. Esto, debe conducirnos a examinar nuestro problema no en términos de cómo moderar sin afectar a la libertad de expresión, sino de cómo hacerlo de tal modo que se coadyuve a salvaguardar, en la medida de lo posible, esta última.

Así las cosas, la moderación debe concebirse como una actividad cuyas funciones vayan más allá de las de control y organización de contenidos. En efecto, debe plantearse como un instrumento que sirva, además, a la tarea de preservar la libertad de expresión de los usuarios en Internet. Tal planteamiento, sin lugar a dudas, resulta coherente con la doble naturaleza que tienen las plataformas; por un lado, como compañías privadas, y, por otro, como espacios o infraestructuras públicas esenciales para la participación social. Para hacer de la moderación una actividad que, estructural y sistemáticamente, sea capaz de impedir o, en todo caso, prevenir la vulneración de la libertad de expresión es necesario adoptar diferentes medidas y acciones. Sin pretender realizar una enumeración exhaustiva de todas ellas, a continuación, expondremos de forma somera algunas de las que son, a todas luces, las soluciones más relevantes para mejorar la moderación de contenidos desde la perspectiva de la defensa de la libertad de expresión del usuario. En la medida en que los riesgos y amenazas que afloran afectan tanto a la moderación como a la regulación de contenidos, hemos considerado pertinente plantear las distintas propuestas y soluciones en atención a estos dos ámbitos.

4.2. Propuestas y soluciones en el ámbito de la moderación de contenidos

1. Transparencia y rendición de cuentas

Terminar con la opacidad que envuelve a la moderación de contenidos resulta esencial para poder abordar mejor este problema, así como para hacer de aquella una práctica menos restrictiva con la libertad de expresión de los usuarios. Por ello, es fundamental que las plataformas adopten enfoques mucho más transparentes en todas las etapas de sus operaciones. Así, en términos de accesibilidad, es conveniente que las reglas internas de las plataformas sean fácilmente visibles para los usuarios y que estén redactadas de forma clara y comprensible. En este sentido, las plataformas podrían mejorar la comprensión de estas reglas, adjuntando directrices más detalladas que acompañasen la lectura de sus reglas de moderación. Además, resultaría conveniente, a efectos de erradicar la opacidad que rodea a la moderación, que las compañías ofrecieran información acerca de cómo se elaboran las reglas internas, qué medidas se adoptan al detectar un contenido ilícito (i.e., filtrado, bloqueo, retirada, "des-priorización", "des-monetización", etc.), o sobre qué base se eligen los *trusted flaggers* (Sander, 2020). En lo concerniente a la arquitectura opaca que caracteriza a los sistemas algorítmicos, sería adecuado que las plataformas explicaran cuando se utilizan algoritmos, cómo realizan la toma de decisiones o bajo qué criterios y bases de datos actúan.

Pero, la cuestión no es plantear esta exigencia como un mero requisito de carácter formal que deban observar las empresas, sino como un imperativo que sea útil y contributivo a efectos de

mejorar el funcionamiento de la moderación. En este sentido, es fundamental que las plataformas sean transparentes tanto desde una perspectiva *ex ante*, facilitando información a los usuarios sobre sus reglas internas y procesos de moderación; como desde una perspectiva *ex post*, aportando toda aquella información que sea relevante para el individuo cuyo contenido ha sido eliminado. Por ejemplo, especificando las reglas y motivos por los cuales su contenido ha sido retirado, cómo ha sido identificado, quien ha sido el responsable de la retirada, o, simplemente, qué medios dispone el usuario para apelar la decisión. Asimismo, resultaría aclarativo además de pedagógico, que las plataformas regularmente publicaran ejemplos sobre cómo resuelven casos particularmente complejos o controvertidos.

2. Implantación de medidas de vigilancia y supervisión humana

Dadas las limitaciones que presentan los sistemas algorítmicos y el consiguiente riesgo de que produzcan falsos positivos, la supervisión y vigilancia humana deviene inevitable. Por ello, tanto la adopción de medidas proactivas como las obligaciones impuestas a las plataformas por tribunales o autoridades públicas que estén basadas en el uso de sistemas de filtrado automático, deberían plantearse única y exclusivamente en aquellos casos en los que se pueda garantizar que la toma de decisiones por parte de los algoritmos es posteriormente supervisada por moderadores humanos. Obviamente, esta exigencia no implica que en la práctica no se vulnere la libertad de expresión, pero sí contribuye de alguna manera a dotar de mayores garantías a los procesos de moderación. Además, y para que lo anterior sea realizable, es necesario que las plataformas incrementen tanto cuantitativa como cualitativamente sus equipos de moderadores.

3. Análisis particulares y respuestas proporcionales. La moderación por diseño

Como ya hemos visto, la dificultad a la hora de definir conceptos complejos como "discurso del odio", "terrorismo" o "violencia" puede tener un grave impacto para la libertad de expresión, máxime cuando la moderación es implantada a gran escala y se aplica un mismo concepto a zonas cultural, social e ideológicamente muy diferentes. Si las plataformas definen estas categorías de forma amplia y vaga, esto puede propiciar la retirada indiscriminada de contenidos. Pero, por el contrario, si realizan definiciones demasiado restringidas, corren el riesgo de convertir sus espacios en ambientes hostiles (Sander, 2020). A pesar de ser una cuestión compleja es necesario que las plataformas definan tales conceptos de la manera más clara y concisa posible, y que, a la hora de aplicarlas al caso concreto, se realice un análisis más particular, intentado tener en cuenta la intención y el contexto en el que se publica el contenido. Para lograr esto, sería conveniente establecer redes de colaboración con *stakeholders* locales (académicos, empresas, organismos públicos, usuarios, etc.) que asistan a las plataformas en el desarrollo y aplicación de las normas de moderación.

Por otro lado, la respuesta por parte de las plataformas debe ser adecuada y proporcionada a la infracción cometida. En este sentido, una vez identificado el contenido ilícito, las plataformas deben aplicar la medida menos invasiva o restrictiva para el usuario, dependiendo tanto de sus capacidades y recursos como de la tipología del contenido. Más allá de este aspecto, algunas plataformas están permitiendo cada vez más que los usuarios controlen el tipo de contenido que

visualizan. Esto, sirve a dos funciones: por un lado, los usuarios se protegen a sí mismos de potenciales contenidos ofensivos y, por otro lado, permite a las plataformas gestionar de una manera más flexible y permisiva el discurso en la web (Sander, 2020).

4.3. Propuestas y soluciones en el ámbito de la regulación de contenidos

En lo concerniente a la regulación de contenidos y a la salvaguarda de la libertad de expresión, la DCE simplemente establece en su considerando 46 que la retirada de contenidos "habrá de llevarse a cabo respetando el principio de libertad de expresión". Más allá de esta referencia, la DCE carece de cualquier mecanismo de salvaguarda. Tal y como se ha analizado en los epígrafes correspondientes, tanto la regulación de la DCE como las iniciativas de la Comisión y de algunos Estados pueden conllevar ciertos riesgos y amenazas para la libertad de expresión habida cuenta de que las mismas propician que las plataformas retiren de forma abusiva los contenidos de los usuarios, bien para no perder el puerto seguro previsto en el artículo 14 de la DCE, bien para no quedar expuestas a gravosas sanciones económicas. Para intentar corregir esta circunstancia, al mismo tiempo que reforzar el propio marco regulativo con miras a la futura *Digital Services Act*, proponemos algunas medidas a modo de salvaguardas para la libertad de expresión.

1. Eliminar la distinción entre intermediarios activos y pasivos

La distinción entre intermediarios activos y pasivos debería eliminarse. (De Streel et al, 2020; Kuczerawy, 2018). Si bien es cierto que es una extensión interpretativa del considerando 42 de la DCE, mantener tal distinción no tiene sentido en la actualidad porque la misma puede no reflejar adecuadamente la realidad de las plataformas. En este sentido, cada vez más, surgen nuevas compañías cuya naturaleza y funciones les impide calificarse como intermediarios pasivos, por lo que quedarían impedidas de acogerse al puerto seguro. Para resolver este aspecto, bastaría con incluir dentro del marco de la exención a los intermediarios de *hosting* activos.

2. Establecer y definir procedimientos de *notice-and-action*

Uno de los principales defectos de los que adolece la DCE es la ausencia de previsión alguna acerca de cómo implementar los procedimientos de NA. En este sentido, y a los efectos de su posible regulación, estos procedimientos deben diseñarse con el fin de limitar al mínimo las interferencias con la libertad de expresión de los usuarios. Para ello, sería aconsejable, como primer paso a tomar, desarrollar un marco regulativo armonizado a nivel comunitario para evitar que, dependiendo del Estado miembro donde se opere, las plataformas se vean obligadas a actuar de forma distinta, con los perjuicios que esto conlleva en términos de eficacia y eficacia. Así pues, sería conveniente establecer dicha regulación sobre la base de los principios de claridad, concreción y seguridad jurídica. Para ello, no sólo resulta necesario fijar reglas precisas sino también intentar eliminar o concretar aquellos aspectos que puedan generar

controversia e incertidumbre a los operadores, como es el caso, por ejemplo, del concepto de "conocimiento efectivo".

Por otro lado, y a propósito de identificar e incorporar de una mejor forma las garantías en los procedimientos de NA, resulta conveniente separar estos en tres etapas diferentes: notificación, toma de decisión y revisión (De Gregorio, 2019). En relación con la primera etapa, es fundamental desarrollar sistemas de notificación precisos y válidos para aquellos usuarios que deseen notificar la ilicitud de un contenido (*notice providers*). Una vez realizada la notificación, sería pertinente poner en conocimiento del titular de dicho contenido (*content providers*) la queja planteada, a efectos de que pueda alegar cuanto estime oportuno. Tras las alegaciones planteadas por el titular del contenido, corresponde a las plataformas decidir qué medida adoptar respecto al mismo. En tal caso, estaríamos en la segunda etapa, la toma de decisión. En esta fase, resulta fundamental que la plataforma explique de forma motivada las razones por las cuales ha adoptado la decisión de, por ejemplo, retirar un contenido, y, si el responsable de tal decisión es un algoritmo o un moderador humano.

Finalmente, y dentro de la etapa de revisión, debemos establecer mecanismos que permitan a los usuarios apelar las decisiones de retirar o bloquear sus contenidos. En este sentido, un procedimiento de apelación en el seno de la propia plataforma debe ajustarse a las exigencias que todo debido proceso debe observar, esto es, que sea accesible, asequible, expeditivo, efectivo y transparente (Kuczerawy, 2018). Naturalmente, junto a este mecanismo de apelación, debemos situar como última vía para la salvaguarda de la libertad de expresión de los usuarios, la revisión judicial.

3. Las iniciativas públicas no pueden convertirse en elementos de presión

En último lugar, tanto la Comisión como los Estados miembros deben abstenerse de imponer a las plataformas obligaciones de retirar contenidos en breves espacios de tiempo y bajo la amenaza de cuantiosas sanciones dado su efecto contraproducente para la libertad de expresión. En este sentido, debemos evitar crear incentivos que provoquen la retirada de contenidos legales y legítimos. Pues, de lo contrario, estaremos abriendo la puerta a una posible censura privada. Precisamente, según el estudio realizado por De Steel et al, (2020), para algunas plataformas estos incentivos constituyen la principal amenaza de interferencia injustificada con los derechos fundamentales

5. Conclusiones

Cualquier sociedad que se preste a ser considerada plenamente como democrática debe erguirse sobre los pilares de la discusión y la libertad de expresión; a su vez, siempre tan amenazados y en peligro de ser socavados. Con este ensayo, se ha pretendido aportar algo de luz a lo que podemos considerar que es uno de los principales desafíos que enfrenta este mundo global e interconectado: la salvaguarda de la libertad de expresión y la defensa de una verdadera cultura democrática en la sociedad de las grandes plataformas online.

Tras realizar este estudio, la primera conclusión que podemos extraer es que el problema de la moderación y su impacto en la libertad de expresión de los usuarios de Internet no es causa únicamente de las limitaciones de la moderación, sino que también es resultado del propio marco normativo y de la presión que ejercen tanto la Comisión como algunos Estados sobre las plataformas. En segundo lugar, y a propósito de la cuestión que planteamos en la introducción de este trabajo, la única forma de moderar y salvaguardar, al mismo tiempo, la libertad de expresión, es hacer de aquella un instrumento que vaya más allá del control y organización de contenidos, dotándola de sólidos mecanismos y garantías que nos permitan impedir o, en todo caso, prevenir la vulneración de dicha libertad. Soluciones que, como ya hemos señalado, deben pasar por la necesaria colaboración y cooperación entre las plataformas, los Estados y los usuarios.

Debemos ser conscientes, empero, que estas soluciones pueden servir de punto de partida, pero nunca ser un bálsamo para el problema. En efecto, el fenómeno de la moderación es un complejo poliedro con numerosas aristas, en el que la defensa de la libertad de expresión se mueve entre un intrincado equilibrio de factores. De ahí que, las plataformas puedan resistirse a coadyuvar a este objetivo si ven amenazados sus intereses. Sirva esto, precisamente, para hacernos comprender que, siendo la defensa de la libertad de expresión en el ámbito online, en esencia, la misma empresa que en el offline, las particularidades y complejidades que revisten a las plataformas y al fenómeno de la moderación de contenidos hacen de aquella una tarea mucho más ardua y difícil.

Pero, hay una cuestión que aflora como esencial al plantear este estudio y que, como sociedad, debemos dar respuesta: *¿qué es mejor para la democracia y la defensa de la libertad de expresión, tolerar unos pocos contenidos ofensivos e ilegales o, por el contrario, permitir un enfoque hobbesiano de gobierno por parte de las plataformas, en el que cedamos una porción de nuestra libertad de expresión a cambio de navegar por ambientes complacientes y seguros?*

A nuestro parecer, la respuesta debe ser clara. Siempre será preferible y deseable sustentar la solidez y vigor del árbol de la democracia y la libertad de expresión que perjudicarlo talando algunas de sus ramas emponzoñadas.

6. Bibliografía

Ahlet, C., Mardden, C., y Yung, C. (2004) *How 'Liberty' Disappeared from Cyberspace: The Mystery Shopper Tests Internet Content Self-Regulation*.

Angelopoulos, C., Brody, A., Hins, W., Hugenholtz, B., Leerssen, P., Margoni, T., McGonagle, T., van Daalen, O. y Van Hoboken, J. (2015) Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation. *Institute for Information Law (IViR).*

Arroyo, E. (2019) La Responsabilidad de los Intermediarios en Internet ¿Puertos Seguros a Prueba de Futuro?, *Cuadernos de Derecho Transnacional*, Vol. 12, Núm. 1, pp. 808-837.

Aswad, E. M. (2018) The Future of Freedom of Expression Online, *Duke Law & Technology Review*, Vol. 17, Num. 1, pp. 27-70.

Balkin, J.M. (2018) Free Speech in the Algorithmic Society, *University of California Davis*, Vol. 51, pp. 1149-1210.

Barlow, J.P. (1996) Declaración de Independencia del Ciberespacio, *Periférica Internacional. Revista para el análisis de la cultura y el territorio,* 1(10), pp. 241-242.

Brannon, V.C. (2019) Free Speech and the Regulation of Social Media Content*, Congressional Research Service.*

Chelioudakis, E. (2019) *The Glawischnig-Piesczek v Facebook case: Knock, knock. Who's there? Automated filters online*, CiTiP. Disponible en: https://www.law.kuleuven.be/citip/blog/the-glawischnig-piesczek-v-facebook-case-knock-knock-whos-there-automated-filters-online/ (última consulta 27/08/2019)

Citron, D., y Jurecic, Q. (2018) Platform Justice: Content Moderation at an Inflection Point, *Aegis Series Paper*, Num. 1811.

Coche, E. (2018) Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online, *Internet Policy Review,* 7(4).

Comisión Europea (2016) Código de conducta para la lucha contra la incitación ilegal al odio en Internet. Disponible en: https://ec.europa.eu/commission/presscorner/detail/es/IP_19_805 (última consulta el 10/08/2020)

Comisión Europea (2017 a) Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Lucha contra el contenido ilícito en línea. Hacia una mayor responsabilización de las plataformas en línea. COM (2017) 555 final, 28.9.2017.

Comisión Europea (2017 b) Comunicado de prensa, "*EU Internet Forum: a major step forward in curbing terrorist content on the internet",* 08.12.2016, disponible en: https://ec.europa.eu/commission/presscorner/detail/en/IP_16_4328 (última consulta el 10/08/2020)

Comisión Europea (2018) Propuesta de Reglamento del Parlamento Europeo y del Consejo para la prevención de la difusión de contenidos terroristas en línea. COM (2018) 640 final, 12.9.2018.

Comisión Europea (2020) Libro Blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza, COM (2020) 65 final, 19.02.20202.

De Gregorio, G. (2019) Democratising online content moderation: A constitutional framework, *Computer Law & Security Review,* 36.

De Streel, A., Defreyne, E., Jacquemin, H., Ledger, M., y Michel A. (2020) *Online Platform's Moderation of Illegal Content Online*, Study for the committee on Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament.

Flew, T., Martin, F. y Suzor, N. (2019) Internet regulation as media policy: Rethinking the question of digital communication platform governance, *Journal of Digital Media & Policy,* Vol. 10, num. 1, pp. 33-50.

Gillespie, T. (2018 a) Platforms are not Intermediaries, *Georgetown Law Review*, Vol. 2, pp. 198-216.

Gillespie, T. (2018 b) Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press*.

GOV.UK (2019) *Online Harms White Paper*. Disponible en: https://www.gov.uk/government/consultations/online-harms-white-paper#history (última consulta 10/08/2020)

ICF, Grimaldi Studio Legale Y 21c Consultancy (2018) *Overview of the legal framework of notice-and-action procedures in Member States SMART 2016/0039*, Directorate-General for Communications Networks, Content & Technology.

Kaye, D., Cannataci, J. y Ní Aoláin, F. (2018) Mandates of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression; the special rapporteur on the right to privacy and the special rapporteur on the promotion of human rights and fundamental freedoms while countering terrorism, *United Nations Human Rights Council,* Reference: OL OTH 71/2018.

Keats, D. (2018) Extremist Speech, Compelled Conformity, and Censorship Creep, *Notre Dame Law Review,* Volume 93, Issue 3, pp.1035-1071.

Keats, D. y Wittes, B. (2017) The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity, *Fordham Law Review*, Vol. 86, Issue 2, Art. 3, pp. 401-423.

Keller, D. (2018) Internet Platforms: observations on speech, danger, and money. *Aegis Series Paper,* num. 1807.

Keller, D. (2020) *Broad Consequences of a Systemic Duty of Care for Platforms*, Center for Internet and Society. Disponible en: http://cyberlaw.stanford.edu/blog/2020/06/broad-consequences-systemic-duty-care-platforms. (última consulta el 24/07/2020)

Klonick, K. (2018) The New Governors: The People, Rules, and Processes Governing Online Speech, *Harvard Law Review*, Vol. 131, PP. 1599-1670.

Knight Foundation y Gallup (2020) *Free Expression, Harmul Speech and Censorship in a Digital World.* Disponible en: https://knightfoundation.org/wp-content/uploads/2020/06/KnightFoundation_Panel6-Techlash2_rprt_061220-v2_es-1.pdf (última consulta el 25/08/2020)

Kuczerawy, A. (2017) The Power of Positive Thinking Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression, *JIPITEC*, 8, pp. 226-237.

Kuczerawy, A. (2018) Safeguards for freedom of expression in the era of online gatekeeping, *Auteurs & Media, 2018*, Vol. 2017, Issue 3, pp. 292-303.

Lidsky, L. (2011) Public Forum 2.0., *Boston University Law Review,* Vol. 91, pp. 1975-2028.

Llansó, E., Van Hoboken, J., Leersen, P., y Harambam, J. (2020) Artificila Intelligence, Content Moderation, and Freedom of Expression, *Transatlantic Working Group.*

McNamee, J. (2011) The Slide from "Self-Regulation" to Corporate Censorship, *European Digital Rights*.

Pia, A. (2019) Upload-Filters: Bypassing Classical Concepts of Censorship? *JIPITEC,* 10, pp. 57-65.

Ranking Digital Rights Corporate Accountability Index (2019) Disponible en: https://rankingdigitalrights.org/index2019/ (última consulta el 15/08/2020)

Recomendación (UE) 2018/334 de la Comisión de 1 de marzo de 2018 sobre medidas para combatir eficazmente los contenidos ilícitos en línea. *DOUE* L 63/50, 6.3.2018.

Relator Especial de las Naciones Unidas (2018) Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de expresión, A/HRC/38/35.

Riis, T., & Schwemer, S. F. (2019). Leaving the European Safe Harbor, Sailing towards Algorithmic Content Regulation. *Journal of Internet Law*, 22, (7).

Roberts, S.T. (2018) Digital detritus: "Error" and the logic of opacity in social media content moderation, *First Monday,* Vol. 23, num. 3. Disponible en: https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649 (última consulta el 25/08/2020)

Sander, B, (2020) Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, *Fordham International Law Journal*, Vol.43, 4, pp. 939-1006.

Santa Clara Principles on Transparency and Accountability in the Content Moderation (2018) Disponible en: https://santaclaraprinciples.org/. (última consulta el 15/08/2020)

Singh, S. (2019) Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User Generated Content, *New America*.

Skorup, B., y Huddleston, J. (2019) The Erosion of Publisher Liability in American Law, Section 230, and the Future of Online Curation, *Mercatus Working Paper, Mercatus Center at George Mason University.*

Suzor, N.P., Myers, S., Quodling, A., y York, J. (2019) What Do We Mean When Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation, *Intrenational Journal of Communication,* 13, pp. 1527-1543.

Tworek, H., y Leerssen, P. (2019) An Analysis of Germany's NetzDG Law, *Transatlanting Working Group*.

Van Hoboken, J., Quintais, J.P., Port, J., y Van Eijk, N. (2018) *Hosting Intermediary Services and Illegal Content Online: An analysis of the scope of article 14 ECD in light of developments in the online service landscape*, Final Report, A Study prepared for the European Commission DG Communications Networks, Content and Technology.

Yemini, M. (2018) The New Irony of Free Speech, *The Columbia Science & Technology Law Review,* Vol. XX, pp. 119-194.

York, J.C., y Zuckerman (2019) Moderating the Public Sphere. En Jorgensen, F., (Ed.) *Human Rights in the Age of Platforms*, London, England. The MIT Press, pp. 137-162.