IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

www.uoc.edu/idp

ARTÍCULO

Redes sociales y discurso del odio: perspectiva internacional[1]

Göran Rollnert Liern
Universidad de Valencia

Fecha de presentación: febrero de 2020
Fecha de aceptación: junio de 2020
Fecha de publicación: septiembre de 2020

Resumen

El presente trabajo analiza las medidas legislativas que se obligan a adoptar los Estados parte en el Protocolo adicional al Convenio sobre la Ciberdelincuencia de 2003 para homogeneizar la legislación penal sobre la difusión de «material racista y xenófobo» en internet tipificando como delito determinadas conductas contra personas o grupos por razón de su raza, color, ascendencia, origen nacional o étnico o religión, tratando en primer lugar las cuestiones interpretativas que plantea su redacción y que han sido abordadas por el Informe explicativo. A continuación se estudian los problemas que plantea la aplicación de los criterios que resultan del Protocolo al llamado «discurso del odio» en las redes sociales: la intencionalidad de la conducta y de los efectos de la difusión teniendo en cuenta las especificidades de las redes sociales; la publicidad intencional de la conducta, que se deslinda de las comunicaciones privadas no penalizables por la predeterminación del destinatario, con los problemas que se plantean cuando el destinatario es un grupo de personas; y la definición del material racista y xenófobo que propugna, promueve o incita al odio, la discriminación y la violencia, por remisión a los instrumentos internacionales en los que la incitación es un elemento clave de forma que se incorpora como requisito implícito adicional el «riesgo inminente» según el conocido estándar norteamericano. Finalmente, se examina cómo algunos de estos criterios –en particular, la doctrina del «riesgo inminente»– han sido aplicados en una reciente sentencia del Tribunal Europeo de Derechos Humanos de 2018 sobre el discurso del odio en internet.

Palabras clave

ciberdelincuencia, discurso de odio, libertad de expresión, incitación, Tribunal Europeo de Derechos Humanos

1. Trabajo realizado en el marco del proyecto de I+D+i Retos MICINN "Derechos y garantías frente a las decisiones automatizadas en entornos de inteligencia artificial, IoT, big data y robótica" (PID2019-108710RB-I00, 2020-2022).

Social networks and hate speech: an international perspective

Abstract

The work analyses the legislative measures which the Party States in the Additional Protocol at the 2003 Convention on Cybercrime are obliged to adopt in order to harmonise the criminal legislation on the dissemination of "racist and xenophobic material" on the Internet, classifying as criminal certain conduct against individuals or groups for reasons of their race, colour, descent, national or ethnic origin or religion, addressing firstly the interpretative issues which its composition brings forward and which were addressed by the Explanatory Report. Next, there is a study of the problems posed by the application of the criteria resulting from the Protocol to that which is termed "hate speech" in social networks: the intentionality of both the conduct and the effects brought about by the dissemination, taking into account the specifities of the social networks; intentional dissemination to the public in relation to the conduct, which is separate from private communications that are unpunishable due to the predetermination of the recipient, with the problems which arise when the recipient is a group of people; and the definition of the racist and xenophobic material which advocates, promotes or incites hatred, discrimination and violence, by reference to the international instruments in which incitement is a key element so that "imminent danger" is incorporated as an additional implicit requirement in accordance with the recognised North American standard. Finally, there is an examination of how some of these criteria –in particular, the doctrine of "imminent danger"– have been applied in a recent judgement by the European Court of Human Rights in 2018 on Internet hate speech.

Keywords

cybercrime, hate speech, freedom of expression, incitement, European Court of Human Rights

Introducción

La proliferación del discurso del odio en internet y, en particular, en las redes sociales[2] es un fenómeno constatado hasta el punto de que con la introducción de los denominados «delitos de odio» en nuestro Código Penal en 2015 se ha pretendido hacerle frente mediante la agravación de las penas «cuando los hechos se hubieran llevado a cabo a través de un medio de comunicación social, por medio de internet o mediante el uso de tecnologías de la información» (artículo 510.3).

El Preámbulo de la Ley Orgánica 1/2015, de 30 de marzo, de reforma del Código Penal, ha justificado la nueva regulación de la incitación al odio y a la violencia «por la necesidad de atender compromisos internacionales» (apartado I). Ahora bien, esta remisión a la normativa internacional no es una novedad. La jurisprudencia constitucional relativa al discurso del odio[3] viene incorporando a su argumentación los textos internacionales sobre esta materia desde el primer momento, tendencia que ha ido *in crescendo* en los pronunciamientos más recientes[4]. Asimismo, a partir de la STC 112/2016, de 20 de junio, el Tribunal Supremo se ha sumado también a esta progresiva recepción del marco internacional sobre el discurso del odio, en especial en lo que se refiere al enaltecimiento del terrorismo.

En consecuencia, cualquier análisis jurídico del discurso del odio resultaría incompleto si no atendiera al vector interpretativo de la normatividad internacional[5], siendo esta última el objeto del presente trabajo, que adopta un enfoque muy concreto, centrado en las peculiaridades de las redes sociales por ser este el entorno en el que las expresiones de odio han incrementado exponencialmente su presencia en detrimento del espacio físico.

1. El Protocolo Adicional al Convenio sobre Ciberdelincuencia y la interpretación de sus términos

El único instrumento internacional sobre el discurso del odio que contempla específicamente las expresiones de odio en internet es el Protocolo Adicional al Convenio sobre la Ciberdelincuencia relativo a la penalización de actos de índole racista y xenófoba cometidos por medio de sistemas informáticos[6], de 28 de enero de 2003, en vigor desde el 1 de marzo de 2006 (en España desde el 1 de abril de 2015), cuyo objetivo es armonizar la legislación penal sustantiva referente a la lucha contra la propaganda racista y xenófoba, completando así las provisiones del Convenio.

En su virtud, los Estados parte se obligan a tipificar como delito en su derecho interno determinadas conductas cometidas «por medio de un sistema informático» sobre personas o grupos por razón de su raza, color, ascendencia, origen nacional o étnico o religión. Entre estas conductas se encuentran las amenazas de comisión de delitos graves, los insultos, la difusión o puesta a disposición del público de material negacionista o justificador de genocidios o crímenes contra la humanidad y, especialmente, la difusión o puesta a disposición del público de «material racista y xenófobo» (artículos 3-6).

Los términos del Protocolo plantean al menos tres cuestiones interpretativas que han sido abordadas por el Informe explicativo[7] (el Informe en adelante) que, aunque no proporciona una «interpretación autorizada del Protocolo» (pág. 1) –como él mismo reconoce–, sí facilita la aplicación de sus disposiciones:

a) ¿Qué se entiende por «material racista y xenófobo»? Según el Protocolo, «todo material escrito, toda imagen o cualquier otra representación de ideas o teorías, que

2. Miró-Llinares y Gómez-Bellvís (2020), págs. 13-15.
3. SSTC 214/1991, de 11 de noviembre, FJ 3; 176/1995, de 11 de diciembre, FJ 5; 235/2007, de 7 de noviembre, FJ 5; y 177/2015, de 22 de julio, FFJJ 2, c) y 5, y voto particular de A. Asúa Batarrita, apartados 1 y 4.
4. SSTC 112/2016, de 20 de junio, FFJJ 4 y 6; y 35/2020, de 25 de febrero, FJ 2, b) y c), por remisión a la anterior.
5. Rollnert Liern (2019).
6. Para una visión general del Protocolo y una valoración crítica del mismo, Akdeniz (2008).
7. Council of Europe (2003). Todos los entrecomillados posteriores de este apartado hacen referencia a este documento, salvo indicación en contrario.

propugne, promueva o incite al odio, la discriminación o la violencia» (artículo 2.1). Afirma el Informe que se penaliza la difusión de «ideas y teorías» en cualquier formato (escrito, imágenes o cualquier otra representación de «ideas o teorías» almacenable, procesable y trasmisible por medios informáticos), no tanto porque sea «expresión de sentimientos/creencias/aversión» sino porque puede llevar a «cierta conducta» (pág. 3) en la medida que «propugne, promueva o incite al odio, la discriminación o la violencia». Dicho de otra forma, la relevancia penal del material radica, más que en lo que expresa, en las acciones que podría provocar en terceras personas, en su efecto perlocutivo.

b) Con carácter general, el Protocolo anuda la responsabilidad penal a que las conductas sean cometidas «intencionalmente», si bien, como señala el Informe, en ciertos casos se exige una intención específica adicional; así, en la negación o justificación del genocidio o de crímenes contra la humanidad, cabe condicionar la penalización a la presencia de la «intención de incitar» al odio, discriminación o violencia; y en la cooperación y la complicidad (artículo 7), el cooperante o cómplice debe tener también la intención de que el delito sea cometido. Los redactores del Protocolo acordaron que «el significado exacto de "intencionalmente" debería dejarse a la interpretación nacional», añadiendo que no se puede castigar penalmente a nadie por los delitos descritos en el Protocolo si no tienen la intención requerida. El Informe se refiere como ejemplo a los proveedores de servicios de internet afirmando que no serán criminalmente responsables por alojar una web con material racista y xenófobo si no han tenido la intención exigida por el derecho interno en el caso particular, no siéndoles exigible que supervisen o controlen activamente la conducta y los contenidos de sus clientes (págs. 5-7).

c) Respecto a la publicidad de la conducta, exigida por la propia naturaleza de los delitos de «difundir o poner a disposición del público» material racista y xenófobo o material negacionista o justificador de genocidios o crímenes contra la humanidad, el Informe define la difusión como la «divulgación (*dissemination*) activa» y la puesta a disposición como la acción de subir material a internet para uso de terceros, incluyendo la creación o compilación de hiperenlaces que faciliten acceso al mismo.

La referencia al «público» deja claro que las comunicaciones o expresiones privadas transmitidas informáticamente quedan fuera del ámbito del delito y están protegidas por el derecho al respeto de la vida privada y familiar y de su correspondencia (artículo 8.1 CEDH).

La exigencia de publicidad se incorpora también al delito de insultos[8] racistas o xenófobos que serían penalmente atípicos en comunicaciones privadas. Por contra, en las amenazas racistas o xenófobas, el tipo delictivo se extiende también a las realizadas en comunicaciones privadas.

2. La problemática aplicación de los criterios del Protocolo al discurso del odio en las redes sociales

2.1. La intencionalidad de la conducta (y de los efectos de la difusión)

La intencionalidad es uno de los «principios centrales» que, según el Tribunal Penal Internacional para Ruanda, emergen de la jurisprudencia internacional sobre incitación a la discriminación y la violencia[9]. Este principio ha resultado relativizado, no obstante, por la Comisión Europea contra el Racismo y la Intolerancia (2016) en su Recomendación de política general núm. 15 de 2015 que recomienda penalizar el discurso del odio no solo cuando exista «intención de incitar» sino, alternativamente, cuando «pueda razonablemente esperarse que incite a actos de violencia, intimidación, hostilidad o discriminación» por existir «riesgo inminente» de violencia, hostilidad, intimidación o discriminación. Sin embargo, la generalización de la exigencia de intencionalidad en los instrumentos internacionales sobre discurso del odio –con matices en

8. McGonagle, comentando la definición de insulto postulada en el Informe (pág. 7), señala la dificultad para deslindar los insultos de las ideas que «ofenden» o «chocan», que la STEDH *Handyside* (7 de diciembre de 1976) considera amparadas por la libertad de expresión en su apartado 49 (2012, pág. 472).
9. *Sentencia Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze. Case No. ICTR-99-51-T (3 de diciembre de 2003)*, conocida como «*Media case*», apartados 980-1007.

algunos casos[10]– y, sobre todo, su asunción expresa en el Protocolo al describir las conductas típicas no dejan lugar a dudas acerca de su aplicabilidad al discurso del odio en internet.

La intencionalidad es un requisito subjetivo definitorio del discurso del odio, más determinante incluso, para algunos autores, que las características del individuo o grupo destinatario del mensaje[11] y que el propio contenido de la expresión, que no es en sí mismo el factor decisivo[12]. La valoración de esta intención en el discurso del odio *online* presenta dos singularidades a considerar:

a) La intención del sujeto no se limita, en el caso de la difusión de material, a la propia conducta de distribuirlo, hacerlo circular, diseminarlo o ponerlo a disposición de terceros, sino que se extiende a su propia naturaleza incitadora o promocional del odio, la discriminación y la violencia. Dicho de otra forma, el autor debe querer difundir ese material y tener «conocimiento efectivo del contenido del material»[13], compartiendo así el propósito de que el efecto de su difusión sea propugnar, promover o incitar al odio, la discriminación o la violencia. A decir del Informe, «el acto de difundir o poner a disposición es solo criminal si la intención está también dirigida al carácter racista y xenófobo del material» (Council of Europe, 2003, pág. 6). En este sentido debe interpretarse la exigencia expresa de que la conducta sancionable se lleve a cabo «intencionadamente y sin derecho», de tal manera que no será sancionable cuando la difusión esté justificada por principios o intereses legítimos que, conforme a la legislación interna de cada Estado, excluyan la responsabilidad criminal (finalidades de aplicación de la ley o investigación de delitos, motivos académicos o de investigación, u otros, Council of Europe, 2003, pág. 5).

b) La valoración de la intención debe tener en cuenta las especificidades del medio en el que se produce la conducta, internet, y en particular de las redes sociales. La desinhibición y libertad de tono propia de las redes sociales, reforzada por la sensación de anonimato[14] son aspectos significativos para calificar la intención del sujeto. La valoración de la intención de los mensajes tiene que tener en cuenta los «códigos expresivos de las redes sociales» (Boix Palop, 2016, pág. 82) que no siempre están claros y son conocidos por los emisores de los mensajes[15]. La difusión de un mensaje de odio en redes sociales, retuiteando por ejemplo o enlazando o poniendo en circulación determinado material, no siempre implica adhesión al contenido del mismo con intención de incitar o fomentar, sino que puede buscar informar sobre dichos mensajes por considerarlos un asunto de interés público, denunciarlos o criticarlos abiertamente o con expresiones o imágenes de carácter cómico, irónico o satírico (memes), siendo por tanto necesario evaluar el propósito del emisor «a partir del conjunto de sus mensajes y no aisladamente»[16].

2.2. La publicidad (intencional) de la conducta

La publicidad requerida por el Protocolo para definir las conductas penalizables (salvo las amenazas) es objeto de consideración en el Informe, que, por una parte, excluye las comunicaciones o expresiones privadas del ámbito de aplicación del Protocolo y, por otra, para deslindar casuísticamente las comunicaciones privadas de aquellas que deben considerarse difusión de material racista y xenófobo penamente perseguible, identifica como criterio principal «la intención del emisor de que el mensaje en cuestión será recibido solo por el destinatario predeterminado», intención subjetiva que puede establecerse a partir de «factores objetivos» como «el contenido del mensaje, la tecnología usada, las medidas de seguridad aplicadas y el contexto en el que el mensaje es enviado» (Council of Europe, 2003, pág. 6).

Sin embargo, también afirma que, si hay más de un destinatario simultáneo del mensaje, el carácter público o privado dependerá del número de receptores y de la naturaleza de la relación entre emisor y receptor. Según el Informe, el acceso abierto del material a cualquier persona (en un

10. Rollnert Liern (2019), págs. 95-98.
11. Reed (2009), pág. 81.
12. Titley, Keen y Földi (2015), pág. 28.
13. Teruel Lozano (2015), pág. 93.
14. Falxa (2015), págs. 3-4. Véase García González sobre la sensación de usuario anónimo como factor de riesgo (2015, págs. 10-13).
15. Díez Bueso (2018), pág. 10.
16. Ídem.

chat, en un grupo de noticias o en un foro, son los ejemplos que pone el Informe) encajará en la conducta típica de «poner a disposición del público», incluso cuando se requiera contraseña, siempre que la misma se proporcione a cualquiera o al que cumpla ciertos criterios; no obstante, la naturaleza de la relación entre los participantes en la comunicación deberá tenerse en cuenta para determinar si hubo difusión pública o puesta a disposición del público o si, por el contrario, se trató de una comunicación privada penalmente atípica (Council of Europe, 2003, pág. 6).

La aplicación de este criterio no plantea problemas en mensajes accesibles para cualquier usuario indeterminado al tratarse de un ámbito público, y tampoco en comunicaciones privadas con destinario único determinado. La dificultad radica en delimitar lo público y lo privado en mensajes dirigidos a grupos de destinatarios que, por el número de sus integrantes y las relaciones que mantienen entre ellos (desde amistad en el mundo real hasta la simple condición de «amigo de un amigo» en las redes sociales), ya no pueden considerarse estrictamente cerrados sino semipúblicos o semiprivados –como los grupos amplios de WhatsApp o Telegram con muchos participantes[17]– difuminándose así los límites entre la privacidad y la esfera pública. La combinación de los criterios del número de destinatarios y sus relaciones con el emisor será aquí lo decisivo, pero no deja de ser conflictiva. ¿A partir de qué número de destinatarios la comunicación es pública, aunque sea un grupo cerrado? ¿Hay publicidad cuando la comunicación en un grupo cerrado se dirige a pocos destinatarios, pero no hay criterios selectivos de admisión en el círculo privado o, si los hay, no se aplican en la práctica?

Otro aspecto a considerar es si en la valoración de la relevancia penal de la difusión debe atenderse a la publicidad potencial del mensaje en las redes sociales (al difundirse en abierto o a grupos numerosos de destinatarios no determinados selectivamente) o, por el contrario, a la publicidad «efectiva», de manera que no por publicarse un contenido en una red social debe considerarse, por definición, público[18]. Para Tamarit Sumalla, la publicidad no solo se produce cuando el acceso es libre sino «también a través de las redes sociales con acceso restringido a usuarios registrados, siempre que el mensaje pueda ser transmitido a un amplio y relativamente indeterminado número de personas» (2018, págs. 20-21). Por su parte, Teruel Lozano propone tener en cuenta «por un lado, el canal o espacio de difusión (si se trata de un canal o espacio privado o abierto a un público indeterminado), y, por otro, la audiencia potencial que haya podido tener el mensaje», de modo que, cuando la difusión se haya producido en un canal público, lo determinante serán los destinatarios potenciales, mientras que en los canales privados con un grupo numeroso de destinatarios habría que tener en cuenta los destinatarios efectivos de la comunicación (2018, pág. 25).

Finalmente, si el criterio fundamental para diferenciar las comunicaciones privadas no punibles de la difusión pública tipificada penalmente es, según el Informe, la intención subjetiva del emisor de limitar el mensaje a un determinado destinatario, la difusión pública solo será sancionable cuando sea intencional y voluntaria. Por tanto, si la publicidad no ha sido buscada intencionadamente por el emisor porque es el destinatario quien ha difundido un mensaje privado, será este último el responsable penalmente de la difusión (si ha tenido intención de incitar, promover o fomentar el odio, la violencia o la discriminación); y si no lo ha difundido con esta intención, sino para denunciar o criticar las expresiones del emisor, ni uno ni otro podrán ser incriminados por faltar la intencionalidad en el destinatario que lo difunde y la intención de darle publicidad en el emisor originario del mensaje. Otro tanto cabe decir si la publicidad es imputable al emisor pero no es deliberada, sino consecuencia de una configuración errónea de la privacidad de la cuenta o de cualquier otro factor ajeno a su voluntad consciente.

La falta de plena conciencia de la publicidad de la conducta es otro elemento a ponderar para apreciar la intencionalidad de la publicidad, influyendo en ello la inmaterialidad del medio de difusión o, dicho de otra manera, la desconexión entre la emisión del mensaje en un entorno físico privado y su repercusión en la esfera pública virtual en la que despliega sus consecuencias. Como dice Rodríguez-Izquierdo Serrano, «al poder hacerlo sin salir físicamente del ámbito privado desde el que escribe, donde tenga su terminal, el receptor-emisor no adquiere una conciencia nítida de estar actuando en un espacio público de comu-

17. Boix Palop (2016), pág. 58.
18. Ídem, pág. 89.

nicación» (2017, pág. 140); de la misma forma, esa desconexión «desvirtúa (...) la percepción clara de los límites a la expresión en internet» y del «carácter lesivo» de las conductas que se realizan en el ciberespacio[19].

Ello nos lleva a la cuestión de la repercusión de las características de la comunicación en redes sociales sobre los elementos típicos de las conductas penalizables según el Protocolo, dificultando y complicando la valoración de la intencionalidad y la publicidad. Así, la facilidad de la expresión espontánea en las redes sociales ofrece a personalidades con rasgos de impulsividad y precipitación una forma rápida y sencilla de comunicar mensajes negativos que, en ocasiones, puede tener más que ver con la construcción de una identidad digital propia ante el círculo de contactos o seguidores que con la intención consciente de provocar en los posibles destinatarios la voluntad de realizar actos de odio, violencia o discriminación. Sin embargo, esta facilidad es un arma de doble filo dado que un tribunal puede deducir la intencionalidad de la conducta del hecho de que, teniendo la posibilidad de corregir, rectificar o aclarar el mensaje fácilmente en la misma red, no se haya hecho uso de esa posibilidad.

Por otra parte, la sensación de anonimato e impunidad[20] maximiza la posibilidad de dar rienda suelta a discursos destructivos sin tener que afrontar las consecuencias que tendrían idénticas acciones en la vida real[21] y, al mismo tiempo, sin ser plenamente conscientes de la difusión y trascendencia de sus mensajes.

La propia percepción del entorno de la red social por los usuarios afecta a la publicidad intencional de la conducta. ¿Hasta qué punto los usuarios tienen conciencia de la posible repercusión pública de sus mensajes en las redes o las perciben como prolongación virtual de un espacio de comunicación informal y desenfadado con un círculo limitado de amistades en el que son permisibles expresiones que no harían en un ámbito público?[22] Las redes sociales convierten inmediatamente en públicos actos y comportamientos individuales antes limitados a ambientes y redes personales, existiendo por ello entre los usuarios jóvenes mayor tolerancia a mensajes de odio y violentos en internet que en el entorno real[23]. La falta de conciencia de actuar en un espacio público de comunicación hace que expresiones exaltadas y de odio «que en principio solo se permitiría a sí mismo en un entorno reducido (...) saltan al debate público digital sin que se haya reflexionado sobre la diferencia cuantitativa y cualitativa que le da su difusión en la red»[24].

2.3. El material que propugne, promueva o incite al odio, la discriminación o la violencia y la doctrina del «riesgo inminente»

La definición del material racista y xenófobo como aquel «que propugne, promueva o incite al odio, la discriminación o la violencia» (artículo 2.1) cuya difusión obliga a sancionar penalmente el Protocolo (artículo 3) requiere ser comentada. El Informe parece establecer una escala entre las tres acciones por su menor o mayor efecto sobre la audiencia destinataria: propugnar sería hacer un alegato justificando en abstracto el odio, la discriminación o la violencia; promover implicaría estimularlo o fomentarlo, buscando una influencia más incisiva en la audiencia; e incitar supondría instar o llamar al público al odio, la discriminación o la violencia (Council of Europe, 2003, pág. 3).

El Informe afirma que esta definición «se basa en las definiciones y documentos existentes nacionales e internacionales (ONU, UE) en la medida de lo posible» (pág. 3) por lo que la determinación del alcance y extensión de las acciones sancionables requiere ser contextualizada en el marco de la regulación internacional sobre el «discurso del odio». El Pacto Internacional de Derechos Civiles y Políticos, de 16 de diciembre de 1966 (en vigor desde el 23 de marzo de 1976), es referencia obligada al disponer su artículo 20.2 que «toda apología del odio nacional, racial o religioso que constituya incitación a la discriminación, la hostilidad o la violencia estará prohibida por la ley».

19. Falxa (2015), pág. 3.
20. Jubany y Malin (2015), pág. 16.
21. Gagliardone, I. *et al.* (2015), pág. 8.
22. En este sentido, Boix Palop (2016), pág. 61.
23. Jubany y Malin (2015), págs. 16 y 28.
24. Rodríguez-Izquierdo Serrano (2017), pág. 40.

La noción de incitación constituye así un elemento clave del que no existe interpretación auténtica en el propio Pacto ni en las Observaciones generales y jurisprudencia del Comité de Derechos Humanos. Puede considerarse una interpretación relativamente autorizada[25] la contenida en el Principio 12.1.iii de Camden de 2009 sobre la libertad de expresión y la igualdad: «declaraciones sobre grupos nacionales, raciales o religiosos que puedan crear un riesgo inminente de discriminación, hostilidad o violencia contra las personas que pertenecen a dichos grupos»[26]. Lo decisivo de la incitación es, por tanto, la generación de un riesgo inminente de discriminación, hostilidad o violencia.

Esta definición de la acción de incitar por remisión al estándar de los Principios de Camden incorpora pues el «riesgo inminente» de odio, discriminación o violencia como requisito adicional implícito en la definición del Protocolo. Ello supone asumir el estándar Brandenburg[27] sobre incitación proveniente de Estados Unidos, siendo difícil de aplicar a los mensajes difundidos en redes sociales al implicar una cierta proximidad temporal entre la incitación y el daño producido por el mensaje.

Por este motivo se ha discutido la utilidad del estándar Brandenburg para la incitación en internet[28]: la exigencia del «riesgo inminente» podría cumplirse en el caso de llamadas inmediatas a la acción en redes como Instagram o Snapchat, pero no cuando los efectos se producen a largo plazo[29]. Se ha señalado así la mayor adecuación de la doctrina de las «amenazas verdaderas» (*true threats*[30]) para inspirar la legislación restrictiva del discurso del odio en internet por cuanto no requiere la inminencia del riesgo sino la intención del emisor de provocar temor en el destinatario sin necesidad de producción de un resultado peligroso: «no hay necesidad de demostrar si una persona razonable habría entendido las declaraciones como intimidantes ni de aportar evidencias de la respuesta de la audiencia, sino solamente la intención de amenazar»[31].

Efectivamente, la relación de causalidad entre un mensaje difundido en las redes sociales y la probabilidad de impacto en forma de actos de odio, discriminación o violencia en pocas ocasiones podrá considerarse que cumple con la exigencia de inminencia del riesgo, en particular dada la fugacidad característica de la comunicación en las redes sociales[32]. Pero por muy fugaz que sea un mensaje o un comentario en las redes, sus efectos potenciales pueden desplegarse mucho más allá del momento en que desaparezcan a consecuencia de su permanencia en la red al margen de la voluntad del emisor, en diversos formatos y a través de la itinerancia de los contenidos mediante los hiperenlaces[33], lo que permite hablar de una potencial «indelebilidad» de los mensajes[34]. La arquitectura de las distintas plataformas influye, no obstante, en la probabilidad de que el mensaje en cuestión haga surgir el riesgo de reacciones de hostilidad, discriminación o violencia; así, se ha señalado que mientras Twitter facilita la rápida y potencialmente extensa difusión del mensaje ofreciendo al mismo tiempo la posibilidad de réplicas inmediatas por interlocutores influyentes que pueden neutralizarlo, Facebook permite la coexistencia paralela de múltiples hilos que pueden pasar más inadvertidos pero, al mismo tiempo, permanecer de forma más duradera[35].

25. Rollnert Liern (2019), págs. 84-91.
26. Article 19 (2009).
27. Por referencia a la conocida sentencia *Brandenburg v. Ohio* 395 U.S. 444, 447 (1969), según la cual «las garantías constitucionales de la libertad de expresión y de la libertad de prensa no permiten a un Estado prohibir o proscribir la defensa del uso de la fuerza excepto cuando tal defensa esté dirigida a incitar o producir una inminente ilegalidad y sea probable que incite o produzca tal acción». Al respecto, Rollnert Liern (2014), págs. 253-255.
28. Tsesis (2017), págs. 667-670; y Ring, 2013, págs. 18-19 y 46.
29. Tsesis, ídem.
30. Formulada en las sentencias *Watts v. United States*, 394 U.S. 705 (1969); *Virginia v. Black*, 538 U.S. 343 (2003); y *Elonis v. United States* 135 S. Ct. 2001 (2015).
31. Tsesis (201), pág. 669.
32. Falxa (2014), pág. 6.
33. Titley *et al.* (2015), págs. 13-14.
34. Díez Bueso (2018), págs. 7 y 13; y Boix Palop (2016), pág. 60.
35. Titley *et al.* (2015), págs. 13-14.

Para finalizar con este punto, las especificidades de las dinámicas comunicativas en las redes sociales parecen coexistir mejor con una definición de la incitación elaborada sustancialmente en torno a la intención subjetiva del emisor de provocar una acción en la audiencia que con el requisito de la inminencia temporal del riesgo, sin perjuicio de que no deba prescindirse de parámetros que permitan estimar el impacto de la expresión al menos en términos de probabilidad razonable.

3. El discurso del odio en internet ante el Tribunal Europeo de Derechos Humanos

Aunque el Protocolo solo ha sido utilizado en una ocasión por el TEDH como Derecho europeo e internacional relevante[36], el Tribunal se ha pronunciado recientemente sobre un caso de discurso del odio en internet en el que, a diferencia de tres casos previos sin doctrina significativa a efectos del presente trabajo[37], aplica los criterios tratados en los apartados anteriores.

Se trata de la sentencia de 28 de agosto de 2018 (caso *Savva Terentyev* c. Rusia) que estimó que violaba la libertad de expresión la condena de un ciudadano ruso por incitación al odio por un comentario contra la policía en un blog que, en términos muy insultantes, decía que «sería genial si en el centro de cada ciudad rusa, en la plaza principal (...) hubiera un horno, como en Auschwitz, en el que ceremonialmente todos los días, y mejor aún, dos veces al día (...) policías infieles fueran quemados. La gente los estaría quemando. Este sería el primer paso para limpiar a la sociedad de esta basura policial» (apartado 13).

El criterio de la intencionalidad de la conducta[38] no es aplicado directamente por el Tribunal, aunque sí implícitamente al señalar que los comentarios del recurrente, realizados durante una discusión, mostraban «desaprobación y rechazo emocional» hacia lo que consideraba abusos policiales y que pueden entenderse como una crítica feroz de la situación de la policía en Rusia (apartado 71); dice el Tribunal no estar convencido de que la referencia a la incineración ceremonial de los policías infieles pueda ser interpretada como «una llamada a la exterminación física de los policías por los ciudadanos ordinarios» siendo más bien «una metáfora provocativa que reafirmó frenéticamente el deseo de que la policía se "purificase" de oficiales corruptos y abusadores ("policías infieles") y fue su llamada emocional a tomar medidas para mejorar la situación» (apartado 72), añadiendo que «la destrucción por el fuego en sí misma no puede ser considerada tampoco como una incitación a una acción ilegal, incluida la violencia», puesto que «actos simbólicos de esta clase pueden ser entendidos como una expresión de insatisfacción y protesta más que como una llamada a la violencia» (apartado 74).

Tampoco la intención es mencionada previamente entre los «factores» a tener en cuenta para valorar las injerencias en la libertad de expresión en casos relativos a expresiones incitadoras o justificadoras del odio, la violencia o la intolerancia (resumidos en el caso *Perinçek*), a cuya luz afirma que va a resolver el caso con particular consideración a «la naturaleza y la redacción de las declaraciones impugnadas, el contexto en el que fueron publicadas, su potencial para llevar a consecuencias dañinas y las razones aducidas por los tribunales rusos para justificar la injerencia en cuestión» (apartado 66).

Para valorar el «impacto potencial», el Tribunal parte de la alta publicidad potencial de las publicaciones en internet, señalando que los comentarios enjuiciados fueron publicados en un blog públicamente accesible y que es verdad que el riesgo de daño planteado por los contenidos y comunicaciones en internet para el ejercicio y disfrute de los derechos y libertades es ciertamente mayor que el que supone la prensa, dado que el discurso ilegal, incluyendo el discurso del odio y las llamadas a la violencia, puede ser difundido como nunca antes, por todo el mundo, en

36. En la sentencia *Perinçek* (15 de octubre de 2015), apartado «Instrumentos y materiales relevantes del Consejo de Europa» (apartados 74-76).
37. SSTEDH *Delfi AS* (16 de junio de 2015), apartados 110, 115, 117, 140, 151, 153, 154, 156-159 y 162; *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt* (2 de febrero de 2016), apartado 91; y la Decisión de inadmisibilidad *Pihl* (9 de marzo de 2017), apartados 25 y 37.
38. El recurrente afirmó no haber tenido intención de hacer público el comentario –por ser una respuesta a un comentario anterior– y mucho menos de llamar a ninguna acción contra la policía, por cuanto había usado la exageración provocativa solamente para expresar la idea de que los policías «infieles» debían ser severamente castigados (apartado 17).

cuestión de segundos y en ocasiones permanece de forma persistente disponible en línea (apartado 79).

Afirmación que matiza seguidamente al señalar que, siendo claro que el alcance y potencial impacto de una afirmación lanzada en línea con unos pocos lectores no es ciertamente el mismo que si hubiera sido publicada en páginas web populares y muy visitadas, «es esencial para la valoración de la influencia potencial de una publicación *online* determinar el ámbito de su alcance entre el público» (apartado 79).

Así, después de reprochar a los tribunales rusos que no hayan valorado siquiera si el blog era generalmente muy visitado o el número de usuarios que accedieron, la Corte valora:

1. la publicidad efectiva que tuvo el comentario, publicado en línea durante un mes hasta que el recurrente lo retiró al conocer la existencia del procedimiento penal;

2. aunque el acceso no había estado restringido había atraído muy poca atención pública: algunos de los conocidos del recurrente lo desconocían y fue solo el procedimiento penal lo que suscitó el interés del público;

3. el recurrente no parece ser un bloguero conocido ni un usuario popular de las redes sociales y menos aún un personaje público o influyente cuya condición pudiera haber aumentado el impacto potencial de sus declaraciones.

A partir de estas consideraciones, el Tribunal llega finalmente a la conclusión de que «el potencial del comentario del recurrente para llegar al público y, por ello, para influir en su opinión fue muy limitado» (apartados 80 y 81).

Pero quizá lo más relevante de la sentencia es que aplica el criterio del «riesgo inminente» acogiendo el estándar Brandenburg sobre el «peligro claro e inminente», utilizado con anterioridad en las sentencias *Gül and others* (8 de junio de 2010), apartado 42, y *Kiliç and Eren* (29 de febrero de 2012), apartado 29, pero que por primera vez se proyecta sobre la incitación *online*. Así, señala que las declaraciones no atacaban personalmente a policías identificables sino más bien a la institución (apartado 75), que debería tener un grado particularmente alto de tolerancia al discurso ofensivo «a menos que tal discurso provocativo pueda provocar *inminentes acciones ilegales* con respecto a su personal y exponerlos a un *verdadero riesgo de violencia física*» (cursivas mías), añadiendo que solo en un «contexto muy sensible» de tensión, conflicto armado, lucha contra el terrorismo o disturbios letales en las prisiones, el Tribunal ha entendido que las declaraciones «pudieron alentar una violencia susceptible de poner en riesgo a los miembros de las fuerzas de seguridad» (apartado 77). Considera el Tribunal que ni en las sentencias de los tribunales nacionales ni en las alegaciones del Gobierno se reseñan circunstancias particulares por las que las afirmaciones en cuestión fueran «responsables de producir acciones ilegales inminentes respecto a los policías y de exponerlos a una amenaza real de violencia física»; los tribunales no explicaron por qué la policía, como grupo social, necesitaba protección reforzada ni se refirieron a ningún factor o contexto que pudiera demostrar que los comentarios del recurrente hubieran animado de hecho a la violencia y puesto a la policía en riesgo.

Para los jueces de Estrasburgo, los tribunales rusos se focalizaron en la naturaleza y redacción del comentario, limitándose a la forma y tenor del discurso, sin analizar las declaraciones en el contexto de la discusión y las ideas que trataban de transmitir, sin intentar valorar su potencial para provocar consecuencias lesivas en atención al ambiente social y político, y su alcance, por lo que fallaron en considerar todos los hechos y factores relevantes (apartado 82); así, afirma en el apartado 84 que a pesar de que la redacción de las declaraciones impugnadas era, de hecho, ofensiva, insultante y virulenta (por lo que el solicitante finalmente se disculpó), no puede considerarse que susciten emociones primarias o prejuicios arraigados en un intento de incitar al odio o la violencia contra los policías rusos; (..) fue más bien la reacción emocional del solicitante a lo que vio como un caso de conducta abusiva de las fuerzas de policía.

Finaliza el Tribunal afirmando que no se puede concluir que el comentario en cuestión tuviera capacidad para provocar violencia sobre la policía rusa generando de este modo «un peligro claro e inminente» que exigiese perseguir y condenar al recurrente (apartado 84), dando así carta de naturaleza a la aplicación del estándar norteamericano del peligro inminente al discurso del odio *online*. Habrá que esperar a ver si se consolida

esta doctrina[39] de la inminencia del peligro con la consiguiente dificultad de su aplicación a las redes sociales.

4. Conclusión

La aplicación de los criterios establecidos en el Protocolo a la penalización del discurso del odio en el ciberespacio resulta problemática en las redes sociales por las singularidades de este entorno.

Así, la valoración de la intencionalidad, que en el caso de la difusión de material debe abarcar su carácter racista o xenófobo y su efecto incitador, tiene que tener en cuenta los códigos propios de las redes sociales (espontaneidad, desinhibición, significado del retuiteo) considerando el conjunto de los mensajes.

La publicidad de la conducta excluye la relevancia penal de las comunicaciones privadas (salvo las amenazas) pero, si lo que determina la naturaleza privada o pública de un mensaje es la predeterminación selectiva del destinatario o la indeterminación del mismo, esta distinción se dificulta cuando los destinatarios integran un grupo numeroso y el acceso al grupo es relativamente abierto. Asimismo, se discute si basta la publicidad potencial de los contenidos o se requiere, en entornos semiprivados con numerosos participantes, una publicidad efectiva.

Por otra parte, el emisor solo responderá penalmente si buscó intencionadamente la difusión pública y al valorar esta intencionalidad de la publicidad deben ponderarse elementos como la plena conciencia de la potencial repercusión pública de sus mensajes allende su círculo de contactos o seguidores, y que no respondan a la construcción de un perfil digital sino a la voluntad de provocar la acción de la audiencia.

La remisión del Protocolo a los documentos internacionales para definir el material racista y xenófobo conduce a la noción de incitación presente en el artículo 20.2 PIDCP. La interpretación más extendida de la incitación incorpora un resultado de «riesgo inminente» asumiendo el estándar norteamericano del caso Brandenburg. Siendo este estándar difícilmente aplicable a las redes sociales, por requerir inmediatez temporal entre las declaraciones y el peligro sin tener en cuenta la permanencia y la itinerancia de los mensajes, resulta más apropiada la doctrina de las «amenazas verdaderas», que solo requiere intención intimidatoria.

Reciente jurisprudencia del TEDH, aun sin citar el Protocolo, ha proyectado estos criterios a comentarios ofensivos en un blog, valorando la intencionalidad del autor, su impacto –potencialmente alto al hacerse en internet, pero muy limitado en el caso concreto– y, lo más relevante, aplicando por primera vez el estándar del «peligro claro e inminente» al discurso del odio *online*.

39. En este sentido, Fathaigh y Voorhoof (2019) critican que en la reciente sentencia *Gürbüz and Bayar* (23 de julio de 2019) el TEDH no ha tomado en consideración el «peligro inminente de violencia» como parte del test de incitación a la violencia que sí aplicó en el caso *Savva Terentyev* que comentamos.

Referencias bibliográficas[40]

ARTICLE 19 (2009). *Los Principios de Camden sobre la Libertad de expresión y la Igualdad* [en línea] https://www.article19.org/data/files/pdfs/standards/los-principios-de-camden-sobre-la-libertad-de-expresion-y-la-igualdad.pdf [Fecha de consulta: 11 de junio de 2020].

AKDENIZ, Y. (2008). *An Advocacy Handbook for the Non Governmental Organisations The Council of Europe's Cyber-Crime Convention 2001 and the additional protocol on the criminalisation of acts of a racist or xenophobic nature committed through computer systems* [en línea] https://www.cyber-rights.org/cybercrime/coe_handbook_crcl.pdf [Fecha de consulta: 11 de junio de 2020].

BOIX PALOP, A. (2016). «La construcción de los límites a la libertad de expresión en las redes sociales». *Revista de Estudios Políticos,* núm. 173, págs. 55-112 [en línea] https://doi.org/10.18042/cepc/rep.173.02 [Fecha de consulta: 11 de junio de 2020].

COMISIÓN EUROPEA CONTRA EL RACISMO Y LA INTOLERANCIA (2016). *Recomendación núm. 15 de política general de la ECRI relativa a la lucha contra el discurso del odio y Memorándum explicativo* (8 de diciembre de 2015), CRI (2016) 15 [en línea] https://rm.coe.int/ecri-general-policy-recommendation-n-15-on-combating-hate-speech-adopt/16808b7904 [Fecha de consulta: 11 de junio de 2020].

COUNCIL OF EUROPE (2003). *Explanatory Report to the Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems*. Estrasburgo, 28-1-2003 [en línea] https://rm.coe.int/16800d37ae [Fecha de consulta: 11 de junio de 2020].

DÍEZ BUESO, L. (2018). La libertad de expresión en las redes sociales. En: GONZÁLEZ JIMÉNEZ, A. (coord.). *Implicaciones jurídicas de los usos y comentarios efectuados a través de las redes. IDP. Revista de Internet, Derecho y Política*, núm. 27 [en línea] http://dx.doi.org/10.7238/idp.v0i27.3146 [Fecha de consulta: 11 de junio de 2020].

FALXA, J. (2015). «Redes sociales y discursos de odio: Un enfoque europeo». En: *Moderno discurso penal y nuevas tecnologías: Memorias [del] III Congreso Internacional de Jóvenes Investigadores en Ciencias Penales* [en línea] https://www.academia.edu/8703187/Redes_sociales_y_discursos_de_odio._Un_enfoque_europeo [Fecha de consulta: 11 de junio de 2020].

FATHAIGH, R.; VOORHOOF, D. (2019). «ECtHR engages in dangerous "triple pirouette" to find criminal prosecution for media coverage of PKK statements did not violate Article 10». En: *Strasbourg Observers* [blog] [en línea] https://biblio.ugent.be/publication/8638098 [Fecha de consulta: 11 de junio de 2020].

GAGLIARDONE, I.; GAL, D.; ALVES, T.; MARTÍNEZ, G. (2015). *Countering online hate speech.* UNESCO [en línea] http://egalitecontreracisme.fr/sites/default/files/atoms/files/countering_online_hate_speech_3.pdf [Fecha de consulta: 11 de junio de 2020].

40.En todas las referencias a bibliografía en inglés debe entenderse que la cita es al original siendo la traducción del autor.

GARCÍA GONZÁLEZ, J. (2015). «Oportunidad criminal, internet y redes sociales: Especial referencia a los menores de edad como usuarios más vulnerables» [en línea]. *Indret: Revista para el Análisis del Derecho,* núm. 4 [en línea] https://indret.com/oportunidad-criminal-internet-y-redes-sociales/ [Fecha de consulta: 11 de junio de 2020].

JUBANY, O.; ROIHA, M. (2015). *Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis*. PRISM [en línea] https://doi.org/10.2991/sschd-16.2016.143 [Fecha de consulta: 11 de junio de 2020].

MCGONAGLE, T. (2012). «A survey and critical analysis of Council of Europe strategies for countering "hate speech"». En: HERZ, M.; MOLNAR, P. (eds.). *The content and context of hate speech: rethinking regulation and responses.* Cambridge: Cambridge University Press, págs. 456-498.

MIRÓ-LLINARES, F.; GÓMEZ-BELLVÍS, A. B. (2020). «Freedom of expression in social media and criminalization of hate speech in Spain: Evolution, impact and empirical analysis of normative compliance and selfcensorship». *Spanish Journal of Legislative Studies,* núm. 1, págs. 1-42 [en línea] https://doi.org/10.21134/sjls.v0i1.1837 [Fecha de consulta: 11 de junio de 2020].

REED, C. (2009). «The challenge of hate speech online». *Information & Communications Technology Law Routledge*, 18 (2), págs. 79-82 [en línea] https://doi.org/10.1080/13600830902812202 [Fecha de consulta: 11 de junio de 2020].

RING, C. E. (2015). «Hate speech in social media: an exploration of the problem and its proposed solutions», *Journalism & Mass Communication Graduate Theses & Dissertations,* núm. 15 [en línea] file:///C:/Users/Usuario/Downloads/hateSpeechInSocialMediaAnExplorationOfTheProblemAndIt.pdf [Fecha de consulta: 11 de junio de 2020].

RODRÍGUEZ-IZQUIERDO SERRANO, M. (2017). «Hate speech y sociedad de la información: La difusión del odio en Internet y las redes sociales». En: ALONSO, L.; VÁZQUEZ, V. J.; CORTINA, A. *Sobre la libertad de expresión y el discurso del odio: Textos críticos.* Sevilla: Athenaica, págs. 129-143.

ROLLNERT LIERN, G. (2014). «Incitación al terrorismo y libertad de expresión: el marco internacional de una relación problemática». *Revista de Derecho Político*, núm. 91, págs. 231-262 [en línea] https://doi.org/10.5944/rdp.91.2014.13675 [Fecha de consulta: 11 de junio de 2020].

ROLLNERT LIERN, G. (2019). «El discurso del odio: una lectura crítica de la regulación internacional». *Revista Española de Derecho Constitucional,* núm. 115, págs. 81-109 [en línea] https://doi.org/10.18042/cepc/redc.115.03 [Fecha de consulta: 11 de junio de 2020].

TAMARIT SUMALLA, J. M. (2018). «Los delitos de odio en las redes sociales». En: GONZÁLEZ JIMÉNEZ, A. (coord.). *Implicaciones jurídicas de los usos y comentarios efectuados a través de las redes. IDP. Revista de Internet, Derecho y Política,* núm. 27 [en línea] https://doi.org/10.7238/idp.v0i27.3151 [Fecha de consulta: 11 de junio de 2020].

TERUEL LOZANO, G. M. (2015). *La lucha del derecho contra el negacionismo: una peligrosa frontera. Estudio constitucional de los límites penales a la libertad de expresión en un ordenamiento abierto y personalista*. Madrid: Centro de Estudios Políticos y Constitucionales.

TERUEL LOZANO, G. M. (2018). «Internet, incitación al terrorismo y libertad de expresión en el marco europeo». *Indret. Revista para el Análisis del Derecho,* núm. 3 [en línea] https://indret.com/internet-incitacion-al-terrorismo-y-libertad-de-expresion-en-el-marco-europeo/ [Fecha de consulta: 11 de junio de 2020].

TITLEY, G.; KEEN, E.; FÖLDI, L. (2014). *Starting points for combating hate speech online. Three studies about online hate speech and ways to address it*. Council of Europe [en línea] https://rm.coe.int/starting-points-for-combating-hate-speech-online/16809c85ea [Fecha de consulta: 11 de junio de 2020].

TSESIS, A. (2017). «Terrorist speech on social media». *Vanderbilt Law Review,* núm. 70 (2), págs. 651-708 [en línea] https://ssrn.com/abstract=2782050 [Fecha de consulta: 11 de junio de 2020].

Cita recomendada

ROLLNERT LIERN, Göran (2020). «Redes sociales y discurso del odio: perspectiva internacional», *IDP. Internet, Derecho y Política,* N.º 31, págs. 1-14. UOC [Fecha de consulta: dd/mm/aa]. http://dx.doi.org/10.7238/idp.v0i31.3233

Sobre el autor

Göran Rollnert Liern
goran.rollnert@gmail.com
Universidad de Valencia

Doctor en Derecho y licenciado en Ciencias Políticas, profesor titular de Derecho Constitucional en la Universidad de Valencia. Miembro de la Red DerechoTICS (www.derechotics.com), ha participado en varios proyectos de investigación sobre libertades informativas, gobierno abierto y redes sociales. Autor de tres monografías, 29 capítulos de libro, 24 artículos en revistas académicas y director de un libro colectivo. Sus líneas de investigación principales son la jefatura del Estado en la monarquía parlamentaria y las libertades ideológicas y de expresión en relación con la incitación al terrorismo y el discurso del odio.