ARTÍCULO

LA INTELIGENCIA ARTIFICIAL Y LA MODERACIÓN DE LOS DISCURSOS DE ODIO EN INTERNET

Natalie Alkiviadou

- *¿Una combinación arriesgada?* -

RESUMEN

Las redes sociales utilizan cada vez más la inteligencia artificial para combatir la incitación al odio en Internet. La enorme cantidad de contenido, la velocidad a la que se desarrolla y la creciente presión estatal sobre las empresas para que eliminen rápidamente la incitación al odio de sus plataformas han llevado a una situación delicada. Este texto sostiene que los mecanismos automatizados, que pueden tener conjuntos sesgados de datos y ser incapaces de detectar ciertos matices del lenguaje, no deberían dejar de ser supervisadas en su aplicación a los discursos de odio, pues esto podría dar lugar a violaciones del derecho de expresión y el derecho a la no discriminación.

PALABRAS CLAVE

Libertad de expresión | Discurso de odio | Inteligencia artificial | Plataformas de medios sociales

1 • Introducción

Las Plataformas de Medios Sociales (SMP por el original en inglés o PMS en la traducción al español) son el principal vehículo de comunicación e información. Posibilitan una comunicación sin fronteras y permiten, entre otras cosas, expresiones políticas, ideológicas, culturales y artísticas, dan voz a grupos tradicionalmente silenciados, ofrecen una alternativa a los medios de comunicación dominantes, que pueden estar censurados por el Estado, permiten la divulgación de noticias de actualidad y dan a conocer violaciones de derechos humanos. Sin embargo, como ha señalado Mchangama *et al.*,[1] el uso masivo de PMS proporciona una nueva visibilidad a fenómenos como el odio y el abuso. El uso de PMS ha sido relacionado de modo directo a acontecimientos horribles como el genocidio en Birmania. Consciente de los peligros del discurso violento como un riesgo inminente de violencia, el autor sostiene que debe tenerse cuidado al aceptar la retórica común de que el discurso de odio está extendido por todos los medios sociales, pues estudios empíricos han demostrado lo contrario. Por ejemplo, Siegel *et al.* condujeron un estudio para evaluar si la campaña electoral de Trump de 2016 (y el siguiente periodo de seis meses) dio lugar a un aumento del discurso de odio en Twitter.[2] Con base en un análisis de una muestra de 12,000 millones de tweets, se encontró que entre el 0,001% y el 0,003% de los tweets contenían discurso de odio en todos y cada uno de los días, "una fracción diminuta tanto del lenguaje político como del contenido general producido por usuarios de Twitter estadounidenses."

No obstante, la presión estatal para regular las plataformas con respecto a los discursos de odio está aumentando, lo cual, como se argumenta en este texto, ha llevado al debilitamiento del derecho a la libre expresión y ha contribuido directamente al silenciamiento de grupos minoritarios. El modo como esta nueva realidad está siendo abordada por estados e instituciones, tales como la Unión Europea, es preocupante. Por ejemplo, en 2017, Alemania aprobó la Ley para la Aplicación de la Ley en las Redes Sociales (NetzDG, por su sigla en alemán), que busca combatir discursos ilegales en internet tales como el insulto, la incitación y la difamación religiosa. Por medio de esta ley, las plataformas sociales con un mínimo de 2 millones de usuarios son obligadas a eliminar contenido ilegal, incluyendo discursos de odio y ofensa religiosa, en un plazo de 24 horas o se arriesgan a penas elevadas de hasta 50 millones de euros. Esto se ha convertido en un prototipo de gobernanza de internet en estados autoritarios. En dos informes de Mchangama *et al.*, uno en 2019 y uno en 2020, Justitia constató la adopción de un modelo NetzDG en más de 20 países, varios de los cuales han sido catalogados por la Casa de la Libertad (Freedom House) como "no libres" o "parcialmente libres".[3] Todos los países requieren a las plataformas digitales eliminar categorías poco definidas de contenido que incluyen información falsa, blasfemia/insulto religioso y discurso de odio. Mchangama y Alkiviadou advierten con preocupación que "pocos de estos países tienen las protecciones básicas del estado de derecho y la libertad de expresión contenidas en el precedente alemán."[4] Un modelo similar está siendo utilizado en la actualidad en el ámbito de la Unión Europea (UE) con la Ley de Servicios Digitales (DSA, por su sigla en inglés).[5]

Como respuesta a los requisitos de fortalecer la regulación, debido al riesgo de multas importantes, las plataformas están escogiendo la estrategia de "mejor prevenir que curar" y regulando los contenidos con rigor. No obstante, como ha señalado Llanso, La comunicación en línea de tales plataformas tiene lugar en una escala gigantesca, haciendo imposible que moderadores humanos revisen todo el contenido antes de que se haga disponible. La simple cantidad de contenido en línea también hace del trabajo de revisión, incluso de contenido denunciado, una tarea complicada. Para responder tanto a la necesidad de evitar penas estatales como al aspecto técnico de la escala y cantidad de contenido, las PMS se han basado cada vez más en inteligencia artificial (AI) bajo la forma de mecanismos automatizados que proactiva o reactivamente lidian con contenido problemático, incluyendo los discursos de odio. En resumen, como ha sido subrayado por Dias *et al.*,[7] la IA proporciona a las PMS "herramientas para supervisar un enorme flujo de información en constante crecimiento – que son bien prácticas en la implementación de políticas de contenido." Aunque esto es necesario en áreas relacionadas con, por ejemplo, el abuso infantil y el avance no consentido de actos íntimos entre adultos, el uso de la IA para regular áreas 'grises' de discurso más controvertidas, como el discurso de odio, es compleja. A la luz de estos desarrollos, este artículo examina la utilización de la IA para regular los discursos de odio en las PMS, sosteniendo que los mecanismos automatizados, que pueden tener conjuntos sesgados de datos y ser incapaces de captar los matices del lenguaje, pueden llevar a violaciones de la libertad de expresión y del derecho de no discriminación de grupos minoritarios, silenciando todavía más a grupos ya marginados.

2 • Discurso de odio: nociones y semántica

No existe una definición de discurso de odio aceptada por todo el mundo. La mayoría de Estados e instituciones están adoptando su propia comprensión de lo que implica,[8] sin definirlo.[9] Uno de los pocos documentos, aunque no es vinculante, que han intentado aclarar el significado del término es la Recomendación del Comité de Ministros del Consejo de Europa sobre discursos de odio.[10] El cual declaró que se entenderá que el término "discurso de odio"

> *abarca todas las formas de expresión que difundan, inciten, promuevan o justifiquen el odio racial, la xenofobia, al antisemitismo u otras formas de odio basadas en la intolerancia, incluida la expresión intolerante de nacionalismo agresivo y etnocentrismo, discriminación y hostilidad contra las minorías, migrantes y personas de origen inmigrante.*

El discurso de odio también ha sido mencionado, pero no definido, por el Tribunal Europeo de Derechos Humanos (TEDH). Por ejemplo, considera que el discurso de odio incluye "todas las formas de expresión que difunden, incitan, promocionan o justifican el odio basado en la intolerancia, incluyendo la intolerancia religiosa."[11] La inclusión de la mera justificación del odio demuestra el bajo umbral para considerar el discurso inaceptable. Además, en sus resoluciones, el TEDH ha sostenido que para ser considerado discurso de odio, no es necesario

que el discurso "recomiende directamente a individuos a cometer actos de odio",[12] puesto que se pueden cometer ataques a personas "insultando, ridiculizando o menospreciando a grupos específicos de la población"[13] y que el "discurso utilizando de modo irresponsable puede no ser digno de protección."[14] En este sentido, el TEDH ha establecido la relación entre el discurso de odio y los efectos negativos que pueden tener en sus víctimas, alegando que incluso el discurso libre de violencia que contiene solo insultos tiene el potencial de causar suficiente daño como para que esté justificado limitar la libertad de expresión.

Además, la Agencia de Derechos Fundamentales de la UE (ADF), ha ofrecido dos formulaciones separadas de discurso de odio; la primera diciendo que se "refiere a la incitación o promoción de odio, discriminación u hostilidad hacia un individuo que está motivada por prejuicios contra esa persona debido a una característica particular."[15] En su informe de 2009 sobre homofobia, la ADF sostuvo que el término discurso de odio, tal y como es utilizando en esa sección particular del informe, "incluye un espectro más amplio de actos verbales incluyendo discursos públicos irrespetuosos."[16] La parte particularmente problemática de esta definición es la referencia general a discurso público irrespetuoso, especialmente considerando que instituciones, tales como el TEDH, extienden la libertad de expresión a ideas que "impacten, ofendan o molesten".[17] Esta es la posición formal del Tribunal, aunque en relación a casos de discurso de odio, como se ha mencionado brevemente antes, ha adoptado un umbral rigurosamente bajo en relación a lo que está dispuesto a aceptar como discurso permisible.

En cuanto a las plataformas en sí mismas, aunque está fuera del alcance de este artículo evaluar todas las orientaciones y normativas para las PMS, contemplamos dos enfoques distintos: Facebook e Instagram, por un lado (ambas propiedad de Plataformas Meta Inc.), y Reddit, por otro. Las primeras[18] formulan su comprensión del discurso del odio en base a tres categorías, la primera siendo el discurso violento y deshumanizador y la segunda, afirmaciones de inferioridad, menosprecio, desprecio y otras formas de 'ofensa' como la repulsión. La tercera categoría incluye afirmaciones pertenecientes a la segregación y exclusión. La lista de características protegidas es amplia, incluyendo aspectos como la raza, etnicidad, afiliación religiosa, casta, orientación sexual y enfermedad grave.[19] Reddit[20] adopta un enfoque más protector del discurso, prohibiendo la incitación a la violencia y la promoción del odio. Las características protegidas que utiliza incluyen la raza, color, religión y embarazo, entre otras. Es destacable que todas las plataformas principales, incluyendo las recién mencionadas, así como Twitter,[21] YouTube,[22] y TikTok,[23] incluyen los motivos de raza y religión en la lista de características protegidas.

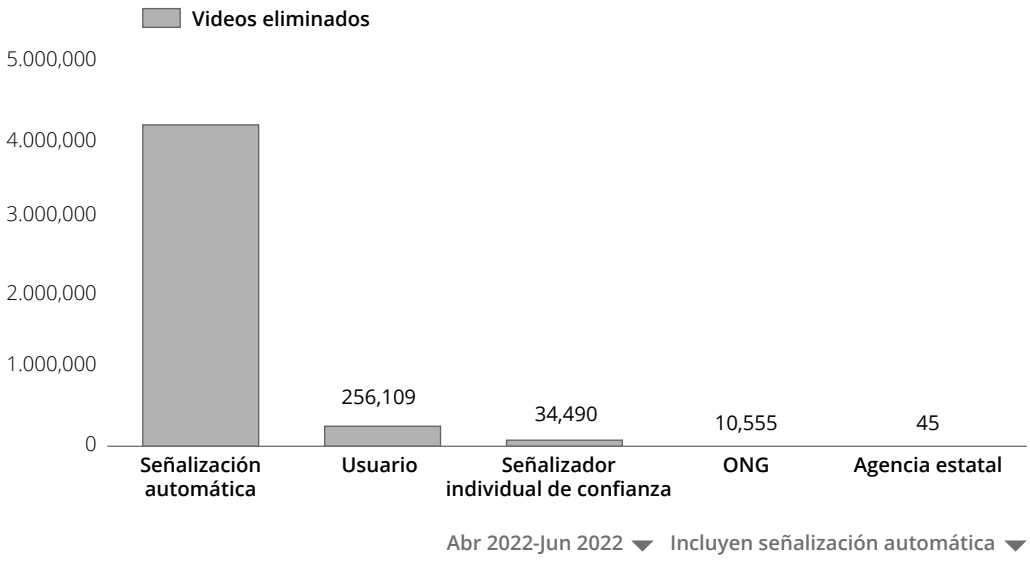
3 • Inteligencia Artificial

La utilización de la IA es una respuesta a la creciente presión del Estado sobre las plataformas sociales para eliminar discursos de odio de un modo rápido y eficaz. Las PMS también son presionadas por otras entidades tales como patrocinadores y usuarios. Para poder cumplir

con tales exigencias (y evitar multas importantes) las empresas utilizan IA, por si sola o en conjunción con moderadores humanos, para eliminar supuestos contenidos de odio. Como señala Dias, tales circunstancias han llevado a las empresas a "actuar proactivamente para poder evitar responsabilidades… en un intento de proteger sus modelos de negocios."[24]

Para ejemplificar la utilización de la IA por las plataformas de medios sociales, uno puede comparar las tasas proactivas de eliminación de discursos de odio entre el primer trimestre de 2018 (de un 38%) y el segundo trimestre de 2022 (de un 95,6%). Como señala un post en el sitio web del Centro de Transparencia (Transparency Center), "nuestra tecnología detecta y elimina proactivamente la gran mayoría de contenido inapropiado antes de que nadie lo denuncie."[25]

En su último informe sobre cumplimiento[26] (segundo trimestre de 2022), YouTube publicó la siguiente ilustración, demostrando el porcentaje de señalización humana y automatizada en todo el espectro de contenido eliminable (no solo discursos de odio):



Dias *et al.* sostienen que los algoritmos desarrollados para conseguir esta automatización son habitualmente creados a medida para tipos de contenido, como fotos, videos, audio y texto.[27] Como Duarte y Llanso descubrieron,[28] las tecnologías actuales detectan texto dañino utilizando procesamiento de lenguaje natural y análisis de sentimientos y, aunque han evolucionado de un modo significativo, su precisión está entre el 70-80%. Sostienen que la IA tiene "capacidad limitada para analizar los matices de significado de la comunicación humana o captar la intención o motivación del hablante." De tal modo que estas tecnologías "todavía no consiguen comprender el contexto, lo que supone riesgos a la libertad de expresión, acceso a la información e igualdad de los

usuarios." Además, Dias *et al.* sostienen que convertir las directrices políticas en códigos de programas puede llevar a cambios de significado, puesto que el lenguaje de máquina está más limitado que su homólogo humano.[29] Debido al poder que las PMS tienen hoy sobre el mercado de la expresión y la información, la creciente necesidad y tendencia a utilizar la IA para lidiar con presiones externas para la eliminación de contenido, así como la cantidad de material, Cowls *et al.* sostienen que hay una necesidad urgente para asegurar que la moderación de contenido tiene lugar de un modo que salvaguarda los derechos humanos y el discurso público.[30]

En consecuencia, y con especial atención en el área polémica del discurso del odio, este artículo examina los riesgos a los derechos humanos que aparecen o pueden aparecer debido al *status quo* actual, concretamente, la creciente dependencia creciente en la IA por parte de empresas privadas con ánimo de lucro, haciendo hincapié en la libertad de expresión y no discriminación.

4 • IA, discursos de odio y los desafíos a la libertad de expresión

El artículo 19 de la Declaración Universal de Derechos Humanos (DUDH) afirma que "[t]odo individuo tiene derecho a la libertad de opinión y de expresión; este derecho incluye el de no ser molestado a causa de sus opiniones, el de investigar y recibir informaciones y opiniones, y el de difundirlas, sin limitación de fronteras, por cualquier medio de expresión."

El derecho a esta libertad también está protegido en otros documentos importantes tales como el Artículo 19 del Pacto Internacional sobre Derechos Civiles y Políticos (ICCPR, por su sigla en inglés) y el Artículo 10 de la Convención Europea de Derechos Humanos. Ambos artículos incluyen limitaciones a la libertad de expresión, mientras que el Artículo 20 de la ICCPR estipula lo siguiente:

> *1 - Toda propaganda en favor de la guerra estará prohibida por la ley.*
> *2 - Toda apología del odio nacional, racial o religioso que constituya incitación a la discriminación, la hostilidad o la violencia estará prohibida por la ley.*

Como señala Dias,[31] basarse en la IA, incluso sin supervisión humana, es necesario cuando se trata de contenido que nunca puede ser justificable ética o legalmente, tal como el abuso infantil. Sin embargo, el tema se complica cuando se trata de áreas de discurso controvertidas, tales como el discurso de odio, para las cuales no hay una postura ética o jurídica universal sobre lo que es y cuando debería ser eliminado (si es que debería serlo en algún caso). En el ámbito de tales discursos, Llanso subraya que el uso de la IA plantea "cuestiones importantes sobre la influencia de la IA en nuestro entorno de información y, en última instancia, sobre nuestros derechos a la libertad

de expresión y acceso a la información".[32] Como señalan Llanso *et al.*,[33] representa "desafíos evidentes a la libertad de expresión y acceso a la información en línea." Un informe del Consejo Europeo señala que la utilización de IA para la regulación del discurso de odio tiene un impacto directo en la libertad de expresión, lo que suscita preocupación por el estado de derecho y, en particular, por las nociones de legalidad, legitimidad y proporcionalidad.[34] El Consejo de Europa señaló que la intensificación del uso de IA para la moderación de contenido puede resultar en un sobre bloqueo y en consecuencia poner en riesgo la libertad de expresión.[35] Gorwa *et al.* afirman que el aumento en el uso de IA amenaza exacerbar la opacidad ya existente en la moderación de contenido, provocar incluso más desconcierto en torno de la justicia en internet y "volviendo a enturbiar la naturaleza fundamentalmente política de las decisiones sobre la expresión que se están ejecutando a gran escala".[36] Además, independientemente de las especificaciones técnicas de un mecanismo en concreto, la identificación (y eliminación) proactiva de discursos de odio constituyen una restricción previa a la expresión, con todas las cuestiones jurídicas que esto implica. Específicamente, Llanso *et al.* sostienen que hay una "gran presunción contra la validez de la censura previa en el derecho internacional de derechos humanos."[37] El antiguo Relator Especial de la ONU sobre la Libertad de Opinión y Expresión, David Kaye, expresó su preocupación acerca del uso de herramientas automatizadas en términos de potenciales sobre bloqueos y sostuvo que las peticiones de ampliar los filtros en la subida de contenidos relacionados con el terrorismo y otras áreas "amenazan establecer regímenes exhaustivos y desproporcionados de censura previa a la publicación."[38]

5 • IA y los desafíos a la no discriminación

Dias sostiene que el uso de la IA puede resultar en una aplicación sesgada de los términos de servicio de las empresas.[39] Esto puede deberse a una falta de datos y/o conjuntos sesgados de datos, dando lugar a un silenciamiento potencial de miembros de comunidades minoritarias.[40] Esto puede llevar a violaciones de la libertad de expresión y el derecho a la no discriminación. En su informe "Mixed Messages: The Limits of Automated Social Content Analysis [Mensajes mezclados: Los Límites del Análisis Automatizado de Contenidos Sociales – traducción libre] el Centro para la Democracia y la Tecnología (Centre for Democracy and Technology) demostró que los mecanismos automatizados pueden impactar el discurso de grupos marginados de un modo desproporcional.[41] Aunque tecnologías como el procesamiento del lenguaje natural y el análisis de sentimientos han sido desarrollados para detectar textos dañinos sin tener que basarse en determinadas palabras o frases, estudios han demostrado, como ya se mencionó antes, que estas tecnologías "todavía están lejos de ser capaces de captar el contexto o detectar la intención o motivación del hablante".[42] Como señala Dias,[43] aunque la comparación de algoritmos se usa mucho para identificar contenido de abuso sexual infantil, no se puede transponer fácilmente a otros casos como el contenido extremista, que "típicamente requiere evaluación del contexto."

En relación a esto, Keller ha observado que la decisión de las plataformas de eliminar contenido islámico extremista afectará "de un modo sistemático e injusto a usuarios inocentes de internet solo por hablar árabe, discutir sobre la política de Oriente Medio o hablar sobre el islam."[44] Hace referencia a la eliminación de una oración (en árabe) subida en Facebook porque supuestamente incumplía las Normas de la Comunidad. La oración decía, "Dios, antes de que acabe este día sagrado, perdona nuestros pecados, bendícenos y a nuestros seres queridos en esta vida y después de muertos con tu misericordia todopoderosa."

Además, como Dias *et al.*,[45] han mostrado, tales tecnologías simplemente no han sido diseñadas para captar el lenguaje de, por ejemplo, la comunidad LGBTQIA+ cuya "burla maleducada" y utilización de términos como "tortillera", "maricón" y "travesti" son una manera de reclamar poder y de preparar a los miembros de esta comunidad a "lidiar con la hostilidad". Dias *et al.* dan varios informes de activistas LGBTQIA+ sobre eliminación de contenido, como la supresión de una mujer trans de Facebook después de mostrar una foto de su nuevo peinado donde se refirió a sí misma como una "travesti".[46] Otro ejemplo utilizado por Dias es un estudio de investigación que mostró que los tweets en inglés afroamericano tienen el doble de posibilidades de ser considerados ofensivos comparados a otros, reflejando la infiltración de prejuicios racistas en la tecnología.[47] Dias *et al.* señalaron los "efectos desconcertantes de los dialectos" que deben ser tomados en cuenta para evitar prejuicios raciales en la detección de discursos de odio.[48] Esto refleja la importancia de contextualizar los discursos, algo que no casa bien con el diseño y ejecución de mecanismos automatizados y que podría plantear riesgos a la participación en línea de grupos minoritarios. Además, los mecanismos automatizados carecen fundamentalmente de la capacidad de comprender los matices y contexto del lenguaje y la comunicación humana. Por ejemplo, YouTube suprimió 6,000 videos documentando el conflicto sirio.[49] Cerró la Agencia de Noticias Qasioun (Qasioun News Agency),[50] un grupo de medios de comunicación independientes informando sobre crímenes de guerra en Siria. Varios videos fueron señalizados como inadecuados por un sistema automático diseñado para identificar contenido extremista. Como señala Dias,[51] otras tecnologías de cotejo algorítmico, tales como PhotoDNA, también parecen operar "ciegos al contexto", que podría ser el motivo por el cual se eliminan esos videos. Facebook prohibió la palabra *kalar* en Birmania, porque los radicales habían dado a esta palabra una "connotación peyorativa" y la habían utilizado para atacar a los Rohingya en Birmania. La palabra fue detectada por mecanismos automatizados que eliminaron posts que la podían haber empleado en otro contexto o con otro significado (incluyendo *kalar oat*, que quiere decir camello). Esto llevó a la eliminación de posts condenando los movimientos fundamentalistas en el país. Fue lo que pasó, por ejemplo, con el siguiente post, donde el usuario opinaba que el nacionalismo extremo y el fundamentalismo religioso son factores negativos:

Fuente: Archivo de la autora

A la vista de los ejemplos mencionados, los problemas de utilizar la IA para lidiar con discursos de odio dan lugar no únicamente a un incumplimiento de la libertad de expresión debido a bloqueos excesivos, sino también a violaciones al derecho a la no discriminación.

6 • Conclusiones

El Consejo de Europa ha propuesto 10 recomendaciones que pueden ser adoptadas para proteger los derechos humanos en lo que se refiere al uso de la IA. Incluyen, por ejemplo, el establecimiento de un marco jurídico para llevar a cabo evaluaciones de impacto en los derechos humanos de los sistemas de IA operativos; la evaluación de sistemas de IA a través de consultas públicas; la obligación de miembros del estado a facilitar la implementación de normativas de derechos humanos en las empresas privadas (como las empresas de medios sociales); supervisión transparente e independiente de sistemas de IA que prestan especial atención a grupos afectados desproporcionadamente por la IA, como minorías étnicas y religiosas; debido atención a los derechos humanos, particularmente a la libertad de expresión; la regla de que la IA siempre debe estar bajo el control humanos y los Estados-miembro deben ofrecer un eficaz acceso a la reparación a las víctimas de violaciones de derechos humanos resultantes del modo en que funciona la IA. También hace referencia a la promoción de la familiarización con la IA. En relación a esto último, hay espacio para ofrecer una formación de derechos humanos y capacitación a quienes directa o indirectamente estén relacionados con la aplicación de sistemas de IA.[52]

Estas recomendaciones sin duda son útiles para mejorar el panorama actual de utilización de mecanismos automatizados para responder a discursos de odio en internet. Sin embargo, las

empresas de medios sociales deben prestar especial atención a temas estructurales que puedan aparecer al utilizar tales mecanismos para eliminar discursos de odio. En primer lugar, debe hacerse hincapié en que, como ha observado Llanso,[53] los temas mencionados no pueden ser abordados con una IA más sofisticada. Además, como señalan Perel y Elink-Koren, "el proceso de traducir instrucciones jurídicas en código conlleva inevitablemente ciertas elecciones sobre cómo la ley es interpretada, que pueden estar afectadas por una variedad de consideraciones extrajudiciales, incluyendo los supuestos profesionales conscientes e inconscientes de los desarrolladores de programas, así como diversas incentivos de la empresa privada."[54] Aunque los mecanismos automatizados pueden ayudar a los moderadores humanos a captar potenciales discursos de odio, no deberían ser los únicos responsables de eliminar los discursos de odio. Conjuntos sesgados de datos de formación, la falta de datos pertinentes y la falta de conceptualización del contexto y los matices pueden conducir a decisiones equivocadas, las cuales pueden tener efectos espantosos en la capacidad de grupos minoritarios de funcionar con igualdad en la esfera de internet.

NOTAS

1 • Jacob Mchangama *et al.*, "A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of Platformization," Justitia, noviembre 2021, visitado el 25 de noviembre de 2022, https://futurefreespeech.com/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf.

2 • Alexandra Siegel *et al.*, "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath." Alexandra Siegel, 6 de marzo de 2019, visitado el 5 de enero de 2022, https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_et_al_election_hatespeech_qjps.pdf.

3 • Jacob Mchangama *et al.*, "A Framework of First Reference," noviembre de 2021.

4 • Jacob Mchangama y Natalie Alkiviadou, "The Digital Berlin Wall: How Germany Built a Prototype for Online Censorship." Euractiv, 8 de octubre de 2020, visitado el 4 de enero de 2022, https://www.euractiv.com/section/digital/opinion/the-digital-berlin-wall-how-germany-built-a-prototype-for-online-censorship/?fbclid=IwAR1fRPCtnP5ce_Glx77uaIB1sIS37BqqHdo-SliBiQWkYmGD3y7f8DaPOi4.

5 • "The Digital Services Act: ensuring a safe and accountable online environment," Comisión Europea, 2022, visitado el 17 de octubre de 2022, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#documents.

6 • Emma J. Llansó, "No Amount of AI in Content Moderation Will Solve Filtering's Prior-Restraint Problem," *Big Data & Society* 7, no. 1 (2020).

7 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online," *Sexuality & Culture* 25 (2021): 700-732.

8 • Informe del Grupo de Expertos del Consejo de Europa para el Desarrollo de los Derechos Humanos (2007) Capítulo IV, 123, párr. 4.

9 • Natalie Alkiviadou, "Regulating Hate Speech in the EU," in *Online Hate Speech in the EU: A Discourse Analytical Perspective*, 1era edición, eds. Stavros Assimakopoulos, Fabienne H. Baider, y Sharon

Millar (Springer Cham, 2017).

10 • Recomendación del Comité de Ministros del Consejo de Europa 97 (20) sobre Discursos de Odio.

11 • Gűndűz c. Turquía, Solicitud Nº. 35071/97 (TEDH 4 de diciembre de 2003) párr. 40; Erbakan c. Turquía, Solicitud Nº. 59405/00 (6 de julio de 2006) párr. 56.

12 • Vejdeland y Otros c. Suecia, Solicitud Nº. 1813/07 (TEDH 9 de febrero de 2012) párr. 54.

13 • *Ibid*.

14 • *Ibid*. párr. 55.

15 • Fundamental Rights Agency, "Hate Speech and Hate Crimes against LGBT Persons" (2009) 1.

16 • Fundamental Rights Agency, "Homophobia and Discrimination on Grounds of Sexual Orientation and Gender Identity in the EU Member States: Part II - The Social Situation" (2009) 44.

17 • The Observer y The Guardian c. el Reino Unido, Solicitud Nº. 13585/88 (TEDH 26 de noviembre de 1991) párr. 59.

18 • "Hate Speech," Meta Transparency Center, 2022, visitadoel25deoctubrede2022,https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech.

19 • "Hate Speech," Meta Transparency Center, 2022, visitado el 2 de noviembre de 2022, https://transparency.fb.com/policies/community-standards/hate-speech/#policy-details.

20 • "Promoting Hate Based on Identity or Vulnerability," Reddit, 2020, visitado el 5 de noviembre de 2022, https://www.reddithelp.com/hc/en-us/articles/360045715951.

21 • "Hateful Conduct Policy," Twitter, 2016, visitado el 2 de enero de 2022, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

22 • "Hate Speech Policy," YouTube, 2019, visitado el 2 de enero de 2022, https://support.google.com/youtube/answer/2801939?hl=en.

23 • "Community Guidelines," TikTok, 2022, visitado el 2 de octubre de 2022, https://www.tiktok.com/community-guidelines?lang=en#38.

24 • Thiago Oliva Dias, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression," *Human Rights Law Review* 20, Nº. 4 (2020): 607-640.

25 • "How Technology Detects Violations," Meta Transparency Center, 19 de enero de 2022, visitado el 3 de noviembre de 2022, https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/.

26 • "YouTube Community Guidelines enforcement," YouTube, 2022, visitado el 3 de noviembre de 2022, https://transparencyreport.google.com/youtube-policy/removals.

27 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

28 • Natasha Duarte y Emma J. Llansó, "Mixed Messages? The Limits of Automated Social Media Content Analysis." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81 (2018): 106-106.

29 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

30 • Joch Cowls *et al.*, "Freedom of Expression in the Digital Public Sphere," AI and Platform Governance, 2020, visitado el 25 de noviembre de 2022, https://doi.org/10.5281/zenodo.4292408.

31 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

32 • Emma J. Llansó, "No Amount of AI...," 2020.

33 • Emma Llanso et al., "Artificial Intelligence, Content Moderation and Freedom of Expression." Transatlantic Working Group, 2020, visitado el 23 de noviembre de 2022, https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

34 • "Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications," Council of Europe, DGI (2017) 12, 2017, visitado el 23 de noviembre de 2022, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, 18.

35 • *Ibid*. 21.

36 • Robert Gorwa *et al.*, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big*

Data & Society 7, Nº. 1 (2020).

37 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation...," 2020.

38 • "Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression," OHCHR, 13 de junio de 2018, visitado el 10 de noviembre de 2022, https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

39 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

40 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation...," 2020.

41 • Natasha Duarte and Emma J. Llansó, "Mixed Messages?...," 2018.

42 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

43 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

44 • Daphne Keller, "Internet Platforms: Observations on Speech, Danger and Money," *Hoover Institution's Aegis Paper Series*, no. 1807 (2018).

45 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

46 • *Ibid.*

47 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

48 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

49 • "YouTube 'made wrong call' on Syria videos'," BBC News, 23 de agosto de 2017, visitado en octubre de 2022, https://www.bbc.com/news/technology-41023234.

50 • *Ibid*.

51 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

52 • "Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights," Consejo de Europa, 2019, visitado el 23 de noviembre de 2022, https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

53 • Emma J. Llansó, "No Amount of AI...," 2020.

54 • Maayan Perel y Niva Elkin-Koren, "Accountability in Algorithmic Copyright Enforcement," *19 Stanford Technology Law Review* 473 (2016), visitado el 25 de noviembre de 2022, https://law.stanford.edu/wp-content/uploads/2016/10/Accountability-in-Algorithmic-Copyright-Enforcement.pdf.

NATALIE ALKIVIADOU – *Chipre/Dinamarca*
Natalie Alkiviadou es investigadora Senior en Justitia (Dinamarca). Sus intereses se centran en la libertad de expresión, la extrema derecha, la incitación al odio y los delitos motivados por el odio. Ha publicado tres monografías y diversos artículos. Alkiviadou es miembro del Centro de Derecho de la Sociedad de la Información de la Università degli Studi di Milano.

contacto: *natalie@justitia-int.org*

Recibido en septiembre de 2022.
Original en inglés. Traducido por Sebastián Porrua.