

LA INTELIGENCIA ARTIFICIAL Y LA LIBERTAD DE EXPRESIÓN EN LA ERA DIGITAL: UN ANÁLISIS JURÍDICO SOBRE EL IMPACTO DE LA MODERACIÓN ALGORÍTMICA EN LAS PLATAFORMAS SOCIALES

ARTÍCULO

LCDA. NANETTE M. GONZÁLEZ DÍAZ*

INTRODUCCIÓN.....	407
I. MARCO CONCEPTUAL: LA INTELIGENCIA ARTIFICIAL Y LA LIBERTAD DE EXPRESIÓN.....	409
A. <i>La inteligencia artificial: concepto y alcance</i>	409
B. <i>Principios generales de la libertad de expresión</i>	410
C. <i>La libertad de palabra ampliada: redes sociales</i>	413
II. LA COLISIÓN ENTRE LA IA Y LA LIBERTAD DE EXPRESIÓN EN LAS REDES SOCIALES..	415
A. <i>Moderación algorítmica:</i>	416
III. IMPACTO DE LA IA EN LA LIBERTAD DE EXPRESIÓN: ALGUNOS POSIBLES PROBLEMAS JURÍDICOS RELACIONADOS A LA MODERACIÓN AUTOMATIZADA.....	421
A. <i>Falta de transparencia</i>	421
B. <i>Sesgos algorítmicos</i>	423
C. <i>La autocensura y el debate público</i>	426
IV. LA VISIÓN INTERNACIONAL: PRINCIPIOS DE DERECHOS HUMANOS PARA LA MODERACIÓN DE CONTENIDO EN LAS REDES SOCIALES	430
CONCLUSIÓN.....	431

INTRODUCCIÓN

En una era dominada por los medios digitales, la libertad de expresión ha encontrado nuevas formas de manifestarse, particularmente a través de las plataformas en línea,¹ donde millones de usuarios interactúan diariamente y generan contenido en tiempo real.² Las redes sociales se han consolidado como un canal fundamental para la difusión de

* Oficial Jurídico adscrita al Panel Central de Investigaciones Jurídicas del Tribunal Supremo de Puerto Rico & Profesora del Departamento de Administración de Empresas & Gerencia de la Universidad de Puerto Rico en Río Piedras.

¹ Sarah C. Rivera Blanco, *Es necesario repensar las protecciones a la libertad de expresión, a la luz de la era digital*, 61 REV. D.P. 115, 118 (2021).

² Alexandra Marie Figueroa Miranda, *'Big brother started following you': Las redes sociales, la libertad de expresión y el neocarpeteo*, 55 REV. JUR. UIPR 469, 486 (2021).

información, noticias, entretenimiento y oportunidades económicas en la vida moderna.³ De hecho, el Pew Research Center reportó que, para el año 2022, más del 70% de los residentes en Estados Unidos obtenían, al menos en una parte, sus noticias a través de redes sociales.⁴ Este incremento exponencial en la circulación de información en línea ha generado la necesidad de establecer mecanismos de regulación y moderación de contenido, con el objetivo de prevenir la proliferación de material perjudicial, como los discursos de odio, la desinformación y el contenido ilegal.

En ese contexto, resulta innegable que el Internet está inundado de contenido potencialmente dañino, cuya difusión representa un riesgo significativo para los objetivos legítimos gubernamentales y la sociedad en general. Tal preocupación ha cobrado mayor relevancia a raíz de casos recientes como *González v. Google LLC*,⁵ y *Twitter, Inc. v. Taamneh*,⁶ en los que familiares de víctimas de ataques terroristas discuten la responsabilidad de las plataformas digitales en la promoción de los objetivos criminales de la organización terrorista ISIS.

En este escenario, la inteligencia artificial (en adelante, “IA”) ha sido fundamental en la creación de herramientas destinadas a la moderación de contenido dañino en redes sociales, al facilitar el análisis y la filtración de grandes volúmenes de datos en tiempo real.⁷ Los algoritmos automatizados permiten a las plataformas identificar patrones y gestionar información que podría indicar posibles infracciones a sus normas internas, haciendo la detección de material problemático más rápida y eficaz.⁸ La *United Nations Educational, Scientific and Cultural Organization* (en adelante, “UNESCO”) ha reconocido que la IA representa una tecnología con el potencial de transformar nuestro futuro.⁹ No obstante, esta

3 CONGRESSIONAL RESEARCH SERVICES, *SOCIAL MEDIA ALGORITHMS: CONTENT RECOMMENDATION, MODERATION, AND CONGRESSIONAL CONSIDERATIONS* 1 (2023), https://crsreports.congress.gov/product/pdf/IF/IF12462?__CF_CHL_TK=P_5S6YT4X7UVHHNTSGHAZCUPCJQGACVYQDX_EcKUBQA-1730454543-1.0.1.1-DV9KB-5YKFPtEH_20PI8_MBJJmVbWmVosR_5FZU_XNY.

4 *Id.*

5 Véase *González v. Google*, 2 F.4TH 871, 880-81 (9TH CIR. 2021). (En el 2015, Nohemí González, una joven estadounidense de veintitrés años, murió víctima de un ataque terrorista mientras disfrutaba de una cena con sus amigos en un café, luego que tres terroristas dispararon contra la multitud de comensales. Al día siguiente, la organización terrorista ISIS se atribuyó la responsabilidad mediante una declaración escrita y la publicación un video en YouTube. Como consecuencia de ello, el padre de la joven presentó una demanda contra Google, Twitter y Facebook, en la que, entre otras cosas, alegó que estas plataformas fueron parte esencial e integral del terrorismo de ISIS al proporcionar una herramienta de comunicación que le permite a esta organización cumplir sus objetivos a través de distintos tipos de contenido y al desempeñar un papel activo en el acto de terrorismo que provocó la muerte de la joven.).

6 Véase *Twitter, Inc. v. Taamneh*, 598 U.S. 471, 478-79 (2023). (En el 2017, Abdulkadir Masharipov llevó a cabo un ataque terrorista en Turquía bajo las órdenes de la organización terrorista ISIS contra una multitud de más de 700 personas. Una de las víctimas que falleció fue Nawras Alassaf, un ciudadano jordano que había visitado Estambul con su esposa para celebrar el año nuevo. Los familiares de las víctimas demandaron a Google, Twitter y Facebook y adujeron que durante años estas empresas han permitido conscientemente que ISIS y sus partidarios utilicen las plataformas para crear contenido que difunde miedo a los usuarios y algoritmos de recomendación como herramientas para reclutar, recaudar fondos y difundir propaganda.).

7 Valentine Crosset & Benoit Dupont, *Cognitive assemblages: The entangled nature of algorithmic content moderation*, 9 (2) BIG DATA & SOCIETY (2022).

8 Julia Haas, *Freedom of the Media and Artificial Intelligence*, GLOBAL CONFERENCE FOR MEDIA FREEDOM (16 de noviembre de 2020), https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-ai-ia.aspx?lang=eng.

9 UNESCO, *Hacia una ética de la investigación en inteligencia artificial a escala mundial*, EL CORREO DE LA UNESCO 3 (2018), https://unesdoc.unesco.org/ark:/48223/pf0000265211_spa.

revolución tecnológica plantea desafíos sin precedentes para la libertad de expresión ya que, en efecto, los sistemas automatizados de moderación pueden conducir a la supresión de contenido legítimo, como consecuencia de las políticas de estas empresas y las limitaciones propias de los modelos algorítmicos.¹⁰

Así, aunque la moderación algorítmica se presenta como una herramienta poderosa que permite gestionar el flujo masivo de información,¹¹ particularmente en la moderación de contenido, se impone una interrogante a considerar: ¿cuál es el costo de su implementación en términos de la libertad de expresión?

Este escrito tiene como objetivo examinar cómo la IA ha redefinido el concepto de libertad de expresión, centrándose en el impacto de la moderación automatizada de contenido y en los desafíos que plantea para las democracias modernas. En primer lugar, se analizarán ciertos conceptos generales como la inteligencia artificial, los principios jurídicos de la libertad de expresión, y la colisión entre ambos. Se analizará cómo funciona la moderación algorítmica, las ventajas y desventajas, cómo estas tecnologías ayudan a gestionar grandes volúmenes de información y los riesgos que representa. Sobre esto último, abordaremos los posibles problemas jurídicos que pueden surgir tales como: los sesgos inherentes a los algoritmos, la falta de transparencia en los procesos de moderación, los riesgos de la autocensura, la amplificación algorítmica y las implicaciones para la diversidad de opiniones y el debate público. Finalmente, se examinará brevemente las guías internacionales sobre la libertad de expresión y cómo estas orientan la moderación de contenido en un entorno globalizado.

I. MARCO CONCEPTUAL: LA INTELIGENCIA ARTIFICIAL Y LA LIBERTAD DE EXPRESIÓN

A. La inteligencia artificial: concepto y alcance:

La IA no es un concepto nuevo, pues sus inicios se remontan previo a la década de 1950.¹² En términos generales, la IA se refiere a “un sistema basado en una máquina que puede, para un conjunto determinado de objetivos definidos por humanos, hacer predicciones, recomendaciones o decisiones que influyen en entornos reales o virtuales”.¹³ A través de la utilización de la IA, “es posible desarrollar computadoras y robots ... que sean capaces de comportarse de [forma] que imiten y vayan más allá de las capacidades humanas”.¹⁴ La base de esta tecnología son los algoritmos o códigos informáticos, “diseñados

¹⁰ Geovanni de Gregorio, *Democratising online content moderation: a constitutional framework*, 36 COMPUTER LAW & SECURITY REVIEW (2020).

¹¹ HAAS, *supra* nota 8, en la pág. 2.

¹² ANICETO PÉREZ Y MADRID, TECNOLOGÍA: INTELIGENCIA ARTIFICIAL. *EL RECONOCIMIENTO FACIAL ES UN SUPERPODER: CÓMO TE AFECTARÁ Y POR QUÉ DEBERÍAS CONOCERLO* 17 (2021), <https://e-archivo.uc3m.es/rest/api/core/bitstreams/72be4448-72cf-4638-a6bc-cd3c5541c7f8/content>.

¹³ Lucía Pérez Martínez, *Implicaciones constitucionales de la inteligencia artificial en el sistema penal*, 63 REV. D.P. 47, 48 (2023) (citando a U.S. Departamento of State, *Artificial Intelligence (AI)*, U.S. DEPARTMENT OF STATE, <https://www.state.gov/ARTIFICIAL-INTELLIGENCE> <https://www.state.gov/ARTIFICIAL-INTELLIGENCE>, (última visita 30 de noviembre de 2024) (traducción suplida).

¹⁴ *Id.* en la pág. 48.

y escritos por seres humanos[, los cuales] ejecutan instrucciones para traducir datos en conclusiones, información o productos”.¹⁵ Estos programas pueden ser habilitados para “analizar y contextualizar datos[,] ... proporcionar información o desencadenar acciones automáticas sin la interferencia humana”.¹⁶

Desde sus inicios y hasta el presente, las herramientas y los usos de la IA se han desarrollado y diversificado a pasos agigantados.¹⁷ Las inversiones en IA han dado lugar a avances transformadores que en la actualidad impactan nuestra vida cotidiana y el bienestar social. Por ejemplo, han brindado grandes beneficios a algunas áreas como la medicina, la educación, el derecho y el bienestar público.¹⁸ Sin embargo, en la medida en que se vuelve más frecuente su uso, debemos ser conscientes de que, a pesar de sus grandes ventajas, la IA también puede cometer errores. De hecho, la IA puede tener más probabilidades de cometer errores que los humanos debido a que generalmente se basa en datos incompletos o inexactos.¹⁹ Lo anterior, plantea una serie de desafíos, ya que las tecnologías fundamentadas en la IA aún son relativamente nuevas, lo cual crea preocupación sobre cómo podría afectar asuntos como las leyes o regulaciones existentes.²⁰ En general, la IA está en constante desarrollo, pero, a su vez, está redefiniendo el futuro y, a medida que se siga adoptando esta tecnología en diversas áreas y procesos en la sociedad, se producirán cambios significativos que podrían representar tanto desafíos como oportunidades.

B. Principios generales de la libertad de expresión:

Uno de los derechos que fungen de ancla a nuestro sistema de vida constitucional es la libertad de expresión. Tal protección encuentra su hogar en la Primera Enmienda de la Constitución de los Estados Unidos.²¹ Del mismo modo, la sección 4 del artículo II de la Constitución de Puerto Rico consagra el derecho fundamental a la libertad de expresión y, ²² como corolario la sección 6 del Artículo citado reconoce el derecho de las personas

¹⁵ ONU, *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión*, 73.er Ses., U.N. Doc. A/73/348, en la pág. 4 (29 de agosto de 2018), <https://documents.un.org/doc/undoc/gen/n18/270/45/pdf/n1827045.pdf>.

¹⁶ *Artificial Intelligence (AI) vs. Machine Learning*, COLUMBIA ENGINEERING (2024), <https://ai.engineering.columbia.edu/ai-vs-machine-learning/>, (última visita 30 de noviembre de 2024) (traducción suplida).

¹⁷ Alfonso Martínez Piovanetti, *La inteligencia artificial en los procesos judiciales y en la litigación puertorriqueña*, MICROJURIS (9 de septiembre de 2024), www.aldia.microjuris.com/2024/09/09/la-inteligencia-artificial-en-los-procesos-judiciales-y-en-la-ligacion-puertorriqueña.

¹⁸ Véase, *Id.*; José Carlos Martínez Valdivieso, *La Inteligencia Artificial: un análisis de la personalidad jurídica y su impacto en el ordenamiento jurídico puertorriqueño*, 54 REV. JUR. UIPR 391, 397-98 (2020).

¹⁹ Pérez Martínez, *supra* nota 13, en la pág. 48. (*citando a* Robayet Syed, *So sue me: Who should be held liable when AI makes mistakes?*, MONASH UNIVERSITY (29 de marzo de 2023), <https://lens.monash.edu/@politics-society/2023/03/29/138545/so-sue-me-wholl-be-held-liable-when-ai-makes-mistakes>).

²⁰ *Id.* (*citando a* Taylor Burton, *AI and the Law: How Artificial Intelligence will Reshape the Legal Industry*, PENNSYLVANIA BAR INSTITUTE (19 de abril de 2023), <https://go.pbi.org/blog/ai-and-the-law-how-artificial-intelligence-will-reshape-the-legal-industry>).

²¹ CONST. EE.UU. enm. 1. (Dispone lo siguiente: “Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances”).

²² CONST. PR art. II, § 4. (Establece que: “No se aprobará ley alguna que restrinja la libertad de palabra o de prensa o el derecho del pueblo a reunirse en asamblea pacífica y a pedir al gobierno la reparación de agravios”).

a asociarse y organizarse para cualquier fin lícito.²³ El porqué de la importancia de este derecho ha sido objeto de debate.²⁴ Sin embargo, no importa las diversas interpretaciones sobre la libertad de expresión, “prácticamente la opinión es unánime al efecto de que su propósito principal es proteger la libre discusión sobre los asuntos de gobierno”.²⁵

El ejercicio efectivo de esta libertad presupone la existencia de otra libertad esencial, que es la libertad de pensamiento y conciencia, pues solamente se puede expresar lo que previamente hemos pensado y, a su vez, esa libertad de pensamiento sólo puede ejercerse si está garantizada la libertad de actuar.²⁶ Asimismo, la libertad de palabra “faculta el desarrollo pleno del individuo y [estimula] el libre intercambio y la diversidad de ideas, elementos vitales del proceso democrático”.²⁷ En virtud de ello, el Gobierno está llamado a la más celosa protección de este derecho.²⁸ Ahora, no se trata de un derecho absoluto y, en ciertas circunstancias excepcionales, el Gobierno podrá intervenir para limitarlo. Al analizar las controversias que plantean el derecho a la libertad de expresión, el Tribunal Supremo de Puerto Rico adoptó unos criterios desarrollados por el Tribunal Supremo federal, bajo el palio de la Primera Enmienda. Específicamente, se bifurcó su análisis en dos categorías, a saber: (1) las acciones gubernamentales que reglamentan el contenido de la expresión y (2) las que regulan su tiempo, lugar y manera.²⁹

En cuanto a la regulación del contenido, en nuestra jurisdicción se ha desfavorecido; debido a que la libertad de expresión protege ideas minoritarias, ideologías controversiales e incluso conductas que podrían considerarse hasta reprochables.³⁰ Por ejemplo, en nuestro ordenamiento jurídico, algunas clases de contenidos protegidos por la cláusula constitucional de la libertad de expresión son: la opinión,³¹ la hipérbole retórica,³² y la parodia

23 CONST. PR art. II, § 6. (Indica lo siguiente: “Las personas podrán asociarse y organizarse libremente para cualquier fin lícito, salvo en organizaciones militares o cuasi militares”).

24 JOSÉ JULIÁN ÁLVAREZ GONZÁLEZ, DERECHO CONSTITUCIONAL DE PUERTO RICO Y RELACIONES CONSTITUCIONALES CON LOS ESTADOS UNIDOS 972 (2009). El Profesor citó al Tribunal Supremo en *Noriega v. Hernández Colon* cuando dicha Curia tomó postura sobre el propósito de la libertad de expresión y dispuso lo siguiente:

Tal vez la teoría más influyente, comúnmente asociada con la prédica de Holmes sobre ‘libre intercambio de ideas’ en el ‘mercado de competencia’, pudiera llamarse la teoría del ‘conocimiento’ o ‘mercado’. Esta teoría postula que el libre intercambio de ideas es indispensable para adelantar el conocimiento y lograr la verdad. Otra teoría prominente, que llamaremos la ‘teoría democrática’, mantiene que el libre flujo de información es una condición esencial para una ciudadanía informada y necesaria para un gobierno propio. Una teoría propuesta, que pudiéramos llamar la ‘teoría de la autorrealización’, enfatiza la relación entre la libre expresión y la autorrealización o autodeterminación del individuo.” (citás omitidas).

Noriega v. Hernández Colon 130 DPR 919, 924 n.1 (1992).

25 *Mari Bras v. Casañas*, 96 DPR 15, 22 (1968).

26 *Muñiz v. Admor. Deporte Hípico*, 156 DPR 18, 23 (2002).

27 *Id.* (citando a *Velázquez Pagán v. A.M.A.*, 131 DPR 568, 576 (1992)).

28 *Asoc. de Maestros v. Srio. de Educación*, 156 DPR 754, 768 (2002).

29 *UPR v. Laborde Torres*, 180 DPR 253, 288 (2010).

30 *Figueroa Miranda*, *supra* nota 2, en la pág. 480.

31 *Asoc. Med. Podiátrica v. Romero*, 157 DPR 240, 246 (2002). (La opinión constituye “una expresión sobre cuestiones de interés público que no contienen una connotación fáctica que sea susceptible de ser probada como falsa”. Es decir, “están protegidas las expresiones que no pueden ser razonablemente interpretadas como que expresan hechos reales”).

32 *Id.* (La hipérbole retórica “constituye una expresión alegadamente difamatoria que no es accionable si se utiliza en sentido figurativo, flexible y no necesariamente por su significado literal”).

o sátira política.³³ A su vez, incluso la mentira no es una categoría censurable o prohibida por la libertad de expresión.³⁴ Por tanto, como regla general, toda limitación de una expresión por su contenido se presumirá inconstitucional.³⁵ Así pues, ante una intervención gubernamental con el contenido de una expresión, tal actuación deberá evaluarse a la luz del escrutinio estricto y ha de existir un interés apremiante y necesario para alcanzarlo.³⁶ Es decir, esto es la forma más restrictiva y desfavorecida por el efecto que pudiera tener en el emisor de la expresión.³⁷

Del mismo modo, existen ciertos tipos de contenidos que no se consideran protegidos por la Primera Enmienda, a saber: las expresiones subversivas, la obscenidad, la difamación, las palabras de riña, la pornografía infantil, el fraude, las amenazas verdaderas o graves y cualquier discurso que represente una amenaza grave e inminente que el Gobierno tenga el deber de prevenir.³⁸ De igual forma, en el año 2010 el Tribunal Supremo de Puerto Rico enumeró de forma específica las diferentes restricciones por contenido que pueden ser aplicables en nuestra jurisdicción;³⁹ entre las que se encuentran expresiones: (1) subversivas; (2) difamatorias; (3) invasoras de la intimidad; (4) obscenas, o (5) de naturaleza comercial.⁴⁰ Las expresiones anteriores estarán sujetas a ser interrumpidas e, incluso, condenadas por los gobiernos dada sus respectivas naturalezas.⁴¹

Si bien es cierto que, como regla general, no se favorece la regulación del contenido de la expresión, el Estado sí puede reglamentar el tiempo, el lugar y la manera de la expresión (dónde, cómo y cuándo nos expresamos) bajo ciertas condiciones.⁴² El profesor José Julián Álvarez explica que “la reglamentación de tiempo, lugar y manera es más tolerada” y que “el grado de tolerancia dependerá del lugar [dónde] se produzca la

33 Vigoreaux *Lorenzana v. Quizno's*, 173 DPR 254, 276 (2008). (“[La] parodia es aquella expresión que utiliza algunos elementos de la composición original para crear una obra nueva que, por lo menos en parte, comente o se burle directamente de la obra original”). *Id.* en la pág. 277. (Por su parte, “la sátira utiliza la obra original sin alteración o transformación significativa para criticar otro elemento social, no relacionado necesariamente con la obra original misma ...”). *Id.* en la pág. 275 (citando a *Garib Bazain v. Clavell*, 135 DPR 475, 488 (1994)). (Así pues, “el humor, sea en forma de sátira, parodia, chistes, etc., rinde una función dual al entretener y servir de crítica social simultáneamente [y.] [c]omo tal, amerita una protección especial en nuestra sociedad”).

34 Véase *U.S. v. Álvarez*, 567 US 709, 722-23 (2012). (Allí, el Sr. Álvarez fue acusado criminalmente en virtud de la ley *The Stolen Valor Act*, por haber mentido al presentarse como un infante de la marina retirado con veinticinco años de servicio y alegar que había sido condecorado con la Medalla de Honor del Congreso. Dicho estatuto tipificó como delito reclamar falsamente la recepción de condecoraciones o medallas militares y establecía una pena mayor si se trataba de una Medalla de Honor del Congreso. Sin embargo, el Tribunal Supremo de Estados Unidos declaró inconstitucional la ley y determinó que las declaraciones del demandado fueron hechas no para algún beneficio económico, sino para ganarse el respeto de los demás. Resolvió que “the Government has not demonstrated that false statements generally should constitute a new category of unprotected speech on this basis”. Añadió que “[p]ermitting the Government to decree this speech to be a criminal offense ...would endorse government authority to compile a list of subjects about which false statements are punishable[,] and [t]hat governmental power has no clear limiting principle”).

35 Laborde Torres, 180 DPR, en las págs. 286-87.

36 ÁLVAREZ GONZÁLEZ, *supra* nota 24.

37 Figueroa Miranda, *supra* nota 2, en la pág. 480.

38 Álvarez, 567 U.S., en la pág. 709.

39 Laborde Torres, 180 DPR, en la pág. 290.

40 *Id.*

41 *Id.* en la pág. 289-90.

42 *Id.* en la pág. 291.

expresión”.⁴³ En el caso del conocido *foro público tradicional*, se comprenden las calles, las aceras y los parques,⁴⁴ puesto que son los foros por excelencia para la divulgación y expresión de ideas.⁴⁵ En estos foros no se puede prohibir de manera absoluta el derecho a la libertad de palabra. No obstante, toda reglamentación se sostendrá siempre y cuando sea neutral sobre el contenido de la expresión, responda a un interés gubernamental significativo y deje abierta alternativas amplias para la expresión.⁴⁶ Por otra parte, en ocasiones, el Estado podría pretender intervenir con algún tipo de expresión al promulgar una regulación de “tiempo, lugar y manera”, cuando, en efecto, pretende censurar el contenido de la expresión.⁴⁷ En este caso, nuestro Tribunal Supremo de Puerto Rico no ha favorecido tal censura.

C. La libertad de palabra ampliada: redes sociales:

Las redes sociales y otras plataformas de intercomunicación virtual han implicado un cambio histórico en la forma en que los seres humanos interactuamos. Durante estas últimas décadas, “los foros tradicionales de comunicación que sirven de escenario para la interacción se han extrapolado, en su gran mayoría, al mundo virtual”.⁴⁸ En la era digital, estos foros son los espacios para el intercambio de ideas, opiniones, información y, a su vez, la expresión del descontento o la disidencia política.⁴⁹ Todas estas interacciones están protegidas por la garantía constitucional de la libertad de expresión.

Con el auge del mundo digital y el aumento en los usuarios en la Internet,⁵⁰ el debate entre los límites de lo público y lo privado se ha renovado debido a que, en la actualidad, no existen garantías jurídicas sobre las expresiones emitidas a través de estos foros.⁵¹ La doctrina de la libertad de expresión “es extensa y en ocasiones amorfa, pero queda un asunto que la Corte Suprema debe resolver: si esta doctrina se debe extender a las redes sociales y, en consecuencia, si los usuarios del internet están protegidos por la Primera Enmienda al expresarse en los foros digitales”.⁵² Aunque esa es una controversia que aún es objeto de debate, para propósitos de este escrito, partimos de la propuesta colectiva que propone

43 ÁLVAREZ GONZÁLEZ, *supra* nota 24.

44 Coss y UPR v. CEE, 137 DPR 877, 887 (1995).

45 Pacheco Fraticelli v. Cintrón Antonsanti, 122 DPR 229, 241-42 (1988).

46 Coss, 137 DPR en la pág. 887.

47 Véase Mari Bras v. Casañas, 96 DPR 15, 21 (1968). (Allí, cinco días antes de las elecciones generales de 1964, la Junta Estatal de Elecciones aprobó varias reglas; entre éstas, prohibió “terminantemente el uso de altoparlantes el día de las elecciones en cualquier parte de Puerto Rico”. El Tribunal Supremo de Puerto Rico resolvió que se pretendía fundamentar tal restricción con la mera alegación de que los altoparlantes perturbaban el ambiente y la tranquilidad de las elecciones. No obstante, concluyó que “[a] poco que se examine la situación se comprenderá que en efecto se trata de una tentativa de silenciar la expresión, no por la selección del método que se utiliza para la diseminación, sino por el mensaje que se intenta diseminar”).

48 *In re Hon. Colón Colón*, 197 DPR 728, 741 (2017).

49 Figueroa Miranda, *supra* nota 2, en la pág. 469.

50 *Id.* en la pág. 486. (Como ilustración, en la segunda década del Siglo XXI se registraron 517.75 millones de usuarios en la red social Facebook y, en el 2018, el número de usuarios ascendió a 2.26 mil millones.).

51 Henry Damián Rodríguez Gracia, *Las redes sociales y la Primera Enmienda: censura gubernamental en la era digital*, 91 REV. JUR. UPR 119, 121 (2022).

52 *Id.*

el reconocimiento de las redes sociales como un foro público moderno y así garantizar la libertad expresión de las personas ante una posible intervención del Estado.⁵³

En ese sentido, es importante analizar la decisión del Tribunal Supremo de los Estados Unidos en *Packingham v. North Carolina*, uno de los primeros casos en que se abordó la relación entre la libertad de expresión y el mundo cibernético.⁵⁴ Específicamente, se analizó el alcance de la libertad de expresión en el Internet a la luz de la Primera Enmienda. En este caso, el Sr. Lester Gerard Packingham, exconvicto, impugnó un estatuto penal de Carolina del Norte que convertía en delito grave que los ofensores sexuales registrados accedieran a sitios de internet y redes sociales que menores de edad solían frecuentar, entre estos: Facebook, Twitter e Instagram. El Tribunal Supremo federal, calificó a las redes sociales como la plaza pública moderna -*modern public square*-; esencialmente una versión digital de una calle o parque público.⁵⁵ De este modo, luego de analizar el propósito de la legislación impugnada, concluyó que esta no fue diseñada específicamente para servir a un interés apremiante del Estado en proteger a los niños de la actividad predatoria en internet. Por lo tanto, la Corte Suprema federal declaró inconstitucional el estatuto por contravenir la Primera Enmienda.

En lo pertinente, el juez asociado Anthony Kennedy, el Juez ponente, explicó que el internet en general y las redes sociales en particular:

(1) [S]e han convertido en los espacios más importantes en la sociedad moderna para el discurso público; (2) proporcionan los “espacios más importantes ... para el intercambio de opiniones” hoy en día; (3) representan un espacio revolucionario para el discurso cívico [,] y (4) proporcionan “las principales fuentes para conocer de los acontecimientos actuales ... y para explorar de otro modo los vastos reinos del pensamiento y conocimiento humano”.⁵⁶

En el caso de Puerto Rico, han sido escasas las oportunidades en las cuales nuestro Tribunal Supremo ha resuelto controversias relacionadas a las redes sociales como mecanismos de expresión. La única instancia en la que nuestro más alto foro estuvo en posición de expresarse sobre el tema fue en *In re Hon. Colón Colón*.⁵⁷ Allí, dicha *Curia* se enfrentó a una controversia ética sobre las redes sociales tras unas expresiones del juez municipal Colón Colón del Tribunal de Primera Instancia. Específicamente, el Juez fue acusado de violentar los Cánones de Ética Judicial al utilizar sus redes sociales, en particular Facebook, para burlarse de las personas que recibía en su sala y discutir incidentes que se desarrollaron en ella.⁵⁸ Aunque el Tribunal realizó un breve análisis sobre la evolución y uso de las redes sociales, basó su determinación en los Cánones de Ética Judicial y no en la interpretación del derecho a la libertad de expresión. En cuanto a las redes sociales, el Tribunal dispuso lo siguiente:

53 *Id.* en la pág. 144.

54 Figueroa Miranda, *supra* nota 2, en las págs. 487-88.

55 Véase *Packingham v. North Carolina*, 582 U.S. 98, 107 (2017).

56 Rodríguez Gracia, *supra* nota 51, en la pág. 141 (*citando a Packingham*, 582 U.S en las págs. 1735-37).

57 Véase *In re Hon. Colón Colón*, 197 DPR 728 (2017).

58 *Id.* en las págs. 732-36.

En las últimas décadas, los foros tradicionales de comunicación que sirven de escenario para la interacción interpersonal se han extrapolado, en su gran mayoría, al mundo virtual. Las redes sociales han tenido una particular preeminencia en esta nueva plataforma de comunicación y de intercambio de información. Las ramificaciones inherentes al uso exponencial de este medio novedoso de comunicación suponen una reconstrucción de nociones tradicionales de conceptos, tales como la privacidad y la libertad de expresión.⁵⁹

II. LA COLISIÓN ENTRE LA IA Y LA LIBERTAD DE EXPRESIÓN EN LAS REDES SOCIALES

La intersección entre la IA y la libertad de expresión en las plataformas sociales revela una dinámica compleja y, en ocasiones, conflictiva. Inicialmente, las plataformas sociales fueron diseñadas por personas con la expectativa de que las redes sociales fueran un espacio para albergar y extender la participación ciudadana y la libre expresión y conexión social.⁶⁰ Sin embargo, aunque las plataformas nos dan la oportunidad de hablar y comunicarnos con una gama más amplia de personas, trajo consigo ciertos peligros como, por ejemplo, la pornografía, la obscenidad, la violencia, la ilegalidad, el odio y el abuso.⁶¹ Para combatir esto último en particular, las empresas han recurrido al uso de la IA para aplicar sus respectivas normas de moderación, las cuales los usuarios están obligados aceptar para poder utilizar la plataforma. A su vez, se reservan la discreción en cuanto a la adopción de medidas sobre el contenido y la cuenta.⁶²

La IA, a través de los algoritmos, es esencial para el funcionamiento adecuado de los sistemas de comunicación los cuales son capaces de analizar texto, imágenes y videos para identificar y eliminar contenido que sea ofensivo, dañino, engañoso, ilegal o inapropiado por cualquier otro motivo.⁶³ No obstante, las mismas herramientas de moderación que las plataformas de redes sociales utilizan para detener el acoso, el discurso de odio y censurar la pornografía también puede silenciar el discurso de muchas otras voces que deberían ser escuchadas.⁶⁴ Esto es lo que sucede con ciertos grupos, como las comunidades LGBTQ+, quienes han denunciado que plataformas como *TikTok* les impiden conservar sus vídeos, limitando su capacidad de compartir sus experiencias de vida con otros usuarios.⁶⁵ Según señalaron representantes de la comunidad, la aplicación silenció el sonido de algunos vídeos o los eliminó por completo por considerar que presuntamente violaban las directrices de la red social.⁶⁶

59 *Id.* en la pág. 741.

60 TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 16-17 del PDF (2018), https://www.researchgate.net/publication/327186182-Custodians_of_the_internet_Platforms_content_moderation_and_the_hidden_decisions_that_shape_social_media.

61 *Id.*

62 ONU, *supra* nota 15 en la pág. 11.

63 Margaret Rouse, *Moderación de contenido*, TECHOPEDIA (15 de enero de 2024), <https://www.techopedia.com/es/definicion/moderacion-contenido>.

64 GILLESPIE, *supra* nota 60, 10-11 del PDF.

65 Allyson Waller, *Tiktok users accuse app of deleting trans content*, DAILY DOT (13 de febrero de 2020), <https://www.dailydot.com/irl/tiktok-censoring-transgender-lgbtq/>.

66 *Id.*

El resultado de ello es que las normas que restringen la difusión de ciertos tipos de discursos en línea suelen colisionar con el derecho a la libertad de expresión que, según los parámetros anteriormente discutidos, debería ser protegido en y fuera de línea. El desafío para las plataformas entonces es, identificar exactamente cuándo, cómo y por qué intervenir y dónde trazar una línea entre lo aceptable y lo prohibido dentro de los límites adecuados de la expresión pública.⁶⁷ Estudios señalan cómo los sistemas que están optimizados para la participación del usuario pueden tener el efecto de restringir, amplificar o priorizar contenidos.⁶⁸ Por consiguiente, intervenir o eliminar el contenido publicado en estas plataformas pudiera conducir fácilmente a la censura, supresión de opiniones, distorsión del discurso público y tener un efecto disuasorio de opiniones disidentes o críticas de los usuarios, lo cual afecta la pluralidad y diversidad de voces en el ámbito digital.⁶⁹

A. Moderación algorítmica:

La moderación de contenido en plataformas en línea es un tema complejo que se enfrenta a múltiples desafíos técnicos, éticos y políticos. Sin embargo, a pesar de su importancia, la moderación de contenidos en Internet ha permanecido hasta hace poco en la oscuridad,⁷⁰ aunque ha existido de distintas formas desde que comenzó la comunicación en línea.⁷¹ Esto, pues los usuarios generalmente utilizan las redes sociales sin ser conscientes de los procesos que implican la filtración y el procesamiento de lo que vemos y publicamos en las redes sociales.

Ahora bien, ¿qué es la moderación algorítmica? El Profesor Giovanni De Gregorio, en su artículo *Democrasting Online Content Moderation: A Constitutional Framework*, define la moderación de contenido en línea como: “the screening, evaluation, categorisation, approval or removal/hiding of online content according to relevant communications and publishing policies. It seeks to support and enforce positive communications behaviour online, and to minimise aggression and anti-social behaviour”.⁷² En otras palabras, la moderación de contenido no solo consiste en la capacidad de decidir qué contenido se permite o cuál se elimina, sino que también juega un papel fundamental en las decisiones de diseño que organizan cómo los miembros de una comunidad interactúan entre sí.⁷³ Es decir, los algoritmos pueden tener la capacidad de determinar qué información de la red social

67 GILLESPIE, *supra* nota 60, 10-11 del PDF.

68 Manuel E. Larrondo & Nicolás Mario Grandi, *La libertad de expresión e inteligencia artificial*, ENCUENTRO DERECHO Y TECNOLOGÍA 2020, 4 del PFD (2020), http://sedici.unlp.edu.ar/bitstream/handle/10915/148272/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y [en adelante, “Larrondo & Grandi, *Libertad y expresión*”]; CONGRESSIONAL RESEARCH SERVICES, *supra* nota 3, 12 del PDF.

69 *Id.*

70 Ronald Durán Allimant, *Trabajo digital y “sonambulismo tecnológico”: el caso de la moderación de contenido en Internet*, ATENEA DIGITAL 7-9 del PDF (2024), <https://atheneadigital.net/article/view/v24-n1-duran/3376-pdf-es>.

71 Roberto Gorwa et al., *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, SAGE JOURNAL (2020), <https://journals.sagepub.com/doi/10.1177/2053951719897945>.

72 De Gregorio, *supra* nota 10.

73 *Id.*

ha de presentarnos,⁷⁴ o, incluso, restringir qué información veremos.⁷⁵ El Profesor Giovanni De Gregorio nos señala que las decisiones de moderación de contenido pueden ser llevadas a cabo por humanos, ser completamente automatizadas o una mezcla de ambas.⁷⁶ No obstante, resalta que las actividades como la priorización de contenido, el bloqueo o la eliminación de la lista de la plataforma suelen ser automatizadas.⁷⁷

La ONU reconoce que los algoritmos son la base de la IA.⁷⁸ El término *algoritmo* se define como “códigos informáticos diseñados y escritos por seres humanos que ejecutan instrucciones para traducir datos en conclusiones, información o productos”.⁷⁹ Toda clase de datos que un algoritmo procesa se traduce y arroja un resultado determinado tal como inferencias, sugerencias o predicciones.⁸⁰ Por ejemplo, en algunos casos, un modelo especializado de algoritmo puede simplemente buscar patrones y señales de publicaciones inadmisibles, mientras que otros sopesan esas señales y deciden qué hacer con el contenido.⁸¹

El funcionamiento cotidiano del Internet exige recoger en gran medida cantidad de información con el fin de que la navegación y las actividades que allí se desarrollan puedan ejecutarse de forma óptima.⁸² Con dicho objetivo, la automatización permite procesar grandes cantidades de datos a una velocidad y en una cantidad que los seres humanos no pueden alcanzar por cuestiones diversas, entre éstas, seguridad pública y salud.⁸³

¿Cómo funciona la moderación algorítmica de contenido? Esta implica una gama de técnicas de estadística e informática, que varían en complejidad y eficacia, según el modelo de algoritmo y su entrenamiento de datos.⁸⁴ En términos generales, el fin de los algoritmos es identificar coincidencias, predecir o clasificar contenido como texto, audio, imágenes o videos. Sin embargo, en el artículo *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, los autores nos explican que existen dos tipos de técnicas principalmente utilizadas en la moderación de contenido algorítmico: clasificación (*classification*) o coincidencia (*matching*).⁸⁵

En el caso de los sistemas de coincidencia, las empresas lo utilizan para moderar miles de millones de publicaciones.⁸⁶ Estos se encargan de identificar si un nuevo conteni-

74 Hiram Meléndez Juarbe, *Vacíos jurídicos: derecho y tecnología*, 87 REV. JUR. UPR 1316, 1317 (2018).

75 Sandra Álvaro, *El Poder de los algoritmos: cómo el software formatea la cultura*, CCCBLAB (29 de enero de 2014), <https://lab.cccb.org/es/el-poder-de-los-algoritmos-como-el-software-formatea-la-cultura/>.

76 De Gregorio, *supra* nota 10.

77 *Id.*

78 ONU, *supra* nota 15, en la pág. 4.

79 *Id.*

80 Tomas Apodaca & Natasha Uzcátegui-Liggett, *How Automated Content Moderation Works (Even When It Doesn't)*, THE MARKUP (1 de marzo de 2024), <https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work>.

81 *Id.*

82 Jennifer Cobbe, *Algorithmic censorship by social platforms: power and resistance*, 34 PHILOS TECHNOL 739, 740 (2020).

83 ONU, *supra* nota 15, en las págs. 5-6.

84 Gorwa et al., *supra* nota 71, en la pág. 3.

85 *Id.*

86 Apodaca & Uzcátegui-Liggett, *supra* nota 80.

do es idéntico o similar a un contenido existente en una base de datos.⁸⁷ Se utilizan, por ejemplo, para detectar imágenes duplicadas o contenido plagiado. Primero se comprimen las publicaciones en fragmentos de texto del tamaño de un bocado de forma que los algoritmos puedan procesarla rápidamente;⁸⁸ a esta información comprimida se le conoce como “hash”.⁸⁹ De muchas formas, los *hashes* funcionan como huellas dactilares, y los algoritmos de moderación de contenido buscan identificar en su base de datos esas huellas dactilares existentes para marcar cualquier coincidencia de contenidos subyacentes que ya saben que quieren moderar.⁹⁰

En cuanto a los sistemas de clasificación, implica categorizar nuevo contenido en una de varias categorías predefinidas.⁹¹ Esto se utiliza, por ejemplo, para identificar contenido como *spam*, discursos de odio o contenido inapropiado. Generalmente, los sistemas de clasificación implican el uso de técnicas de aprendizaje automático.⁹² La técnica de aprendizaje automático consiste en enseñar a los modelos a predecir resultados basados en ejemplos etiquetados previamente.⁹³ Es decir, permite que el software aprenda de forma independiente a partir de un conjunto de datos.⁹⁴ Estos pueden realizar acciones como reconocer el contenido de una foto o analizar el texto de una publicación y, a su vez, se pueden adiestrar para identificar contenido gráfico, de modo que puedan entonces determinar si se elimina de la plataforma o se reduce su visibilidad.⁹⁵ Algunos enfoques comunes incluyen: aprendizaje supervisado (los modelos se entrenan con datos etiquetados para aprender a clasificar nuevos datos), métodos basados en palabras (se analizan las palabras del texto para determinar su categoría), y métodos basados en *embeddings* (utilizan representaciones vectoriales de palabras para capturar las relaciones semánticas).⁹⁶

Los autores destacan que, si bien los avances en la IA han mejorado la precisión de estos métodos, tanto los modelos de clasificación como los de coincidencia enfrentan grandes desafíos durante la moderación de contenido. En particular, apuntan que la coincidencia puede ser limitada por la calidad de la base de datos y la capacidad de los algoritmos para detectar cualquier cambio en el contenido.⁹⁷ De la misma forma, señalan que los modelos de clasificación pueden ser complicados debido a la complejidad del lenguaje natural y la subjetividad de las categorías de datos en estos.⁹⁸

87 Gorwa et al., *supra* nota 71, en la pág. 3.

88 Apodaca & Uzcátegui-Liggett, *supra* nota 80.

89 *Id.*

90 *Id.*

91 Gorwa et al., *supra* nota 71, en la pág. 5.

92 *Id.*

93 Vincent E. Molinari, *Algorithmic harm and design defects: the problem with applying design defect liability standards to machine learning products*, 48 SETON HALL J. LEGIS. & PUB. POL'Y 416, 419–20 (2024).

94 *Algoritmo de clasificación: definición y modelos principales*, DATASCIENTEST (27 de febrero de 2024), <https://datascientest.com/es/algoritmo-de-clasificacion>.

95 Apodaca & Uzcátegui-Liggett, *supra* nota 80.

96 Gorwa et al., *supra* nota 71, en la pág. 5.

97 *Id.*

98 *Id.*

i. ¿Qué tipo de contenido juzga el algoritmo?

Según expuesto previamente, la IA, a través de los algoritmos, ayuda a las empresas de redes sociales a moderar (monitorear, clasificar y eliminar) el contenido, según las normas y reglas de la plataforma. En particular, estas herramientas automatizadas intentan abordar cuestiones como, por ejemplo: contenido relacionado a la violencia gráfica, contenido terrorista, transacciones ilegales, discurso tóxico (discurso de odio, acoso e intimidación), contenido sexual, abuso sexual infantil, detección de spam y cuentas falsas, temas de desinformación, propiedad intelectual, noticias falsas, filtros de palabras claves, procesamiento de lenguas naturales (por el cual la naturaleza del contenido se juzga según palabras o imágenes prohibidas), entre otros algoritmos de detección.⁹⁹ Así, sucesivamente cada categoría contiene docenas de subcategorías, definiciones técnicas y enlaces a materiales complementarios.¹⁰⁰

A partir de esas categorías y subcategorías, surge la siguiente pregunta: ¿Qué sucede cuando el algoritmo de la red social identifica que la expresión publicada por un usuario viola las normas de la plataforma? La moderación de contenido consta de varias etapas que incluyen: una advertencia a la comunidad virtual, técnicas para detectar violaciones a las normas y, finalmente, la aplicación de las políticas de la plataforma.¹⁰¹ Una vez identificado el contenido prohibido, o cuando el algoritmo predice que dicho contenido presenta el material a moderar, pueden ocurrir varios resultados.¹⁰² Así, la IA puede “imponer una advertencia, suspensión o desactivación a la cuenta del usuario por infracción de las condiciones de servicio o para bloquear o filtrar sitios web según datos o contenidos de dominios prohibidos”.¹⁰³ Conviene destacar que no existe un método uniforme o general para la moderación de contenido entre las diferentes plataformas de redes sociales.¹⁰⁴ Algunas compañías emplean normas más restrictivas, mientras que otras prometen una mínima o ninguna moderación de contenido.¹⁰⁵

ii. Capacidades y limitaciones

El volumen de datos generados en línea crece de forma exponencial cada segundo. En solo un minuto, en Google se realizan más de tres millones de búsquedas y, a su vez, en Fa-

⁹⁹ Durán Allimant, *supra* nota 71, en la pág. 8 del PDF; Gorwa et al., *supra* nota 72, en las págs. 5-6; Anastasia Kozyreva, et al., *Resolving content moderation dilemmas between free speech and harmful misinformation*, 120 PNAS 7 (2023), <https://www.pnas.org/doi/10.1073/pnas.2210666120>; Felipe J. Schneider & Marian Andrei Rizoiu, *The effectiveness of moderating harmful online content* 120 PNAS 7 (2023), <https://www.pnas.org/doi/full/10.1073/pnas.2307360120>; ONU, *supra* nota 15, en la pág. 8; Mohit Singhal et al., *SoK: content moderation in social media, from guidelines to enforcement, and research to practice*, IEE 8TH EUROPEAN SYMPOSIUM ON SECURITY (2023), <https://ieeexplore.ieee.org/abstract/document/10190527>.

¹⁰⁰ Andrew Marantz, *Why Facebook can't fix itself*, THE NEW YORKER (12 de octubre de 2020), <https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself>.

¹⁰¹ Singhal et al., *supra* nota 99, en la pág. 859.

¹⁰² Gorwa et al., *supra* nota 71, en las págs. 6.

¹⁰³ ONU, *supra* nota 15, en la pág. 8.

¹⁰⁴ Gorwa et al., *supra* nota 71, en las págs. 8-9.

¹⁰⁵ *Id.*

cebook se envían más de treinta millones de mensajes y se visualizan más de dos millones de videos.¹⁰⁶ Por su parte, en Twitter se publican alrededor de cuatrocientos cincuenta mil tuits, en Instagram se comparten más de cuarenta y seis mil fotos, y en YouTube se difunden más de cuatro millones de horas de videos, mientras que en Netflix se consume casi el doble de ese tiempo en contenido de *streaming*.¹⁰⁷ Dado al enorme volumen de datos que genera una persona al interactuar en la red, el uso de la IA constituye un paso necesario y esencial para las plataformas sociales.

Esta tecnología permite alcanzar un procesamiento masivo de información en tiempo real, lo cual sería imposible de gestionar manualmente de forma rápida y eficiente.¹⁰⁸ Entre sus principales ventajas están la capacidad de identificar contenido perjudicial con rapidez, su alcance global y la adaptabilidad de los algoritmos específicos de cada plataforma para gestionar estos datos.

La moderación de contenido implica que los moderadores de las plataformas deberán, en pocos segundos, tomar decisiones evaluando y eliminando el contenido calificado como inapropiado, según las políticas de la empresa.¹⁰⁹ Tal poder de los algoritmos se refleja cuando en los informes de transparencia de las plataformas digitales, se declara que inequívocamente estas herramientas automatizadas reportan más contenido dañino que los humanos y de forma más rápida.¹¹⁰ Por ejemplo, en cuanto al contenido clasificado como terrorista, Facebook aseguró que la implementación de estas herramientas automatizadas mostraron una tasa de éxito de más del 99.7% de contenido moderado.¹¹¹ A finales de 2020, Facebook tenía más de 2,850 millones de usuarios activos y, en el segundo trimestre del 2021, había eliminado más de 7.1 millones de publicaciones etiquetadas como terroristas.¹¹² No obstante, si bien estos números son impresionantes, tales virtudes se contraponen a la falta de comprensión del contexto, los errores en la interpretación del lenguaje y el impacto desproporcionado en ciertos discursos.¹¹³ Para ilustrar, estudios sugieren que, a pesar de los avances en el aprendizaje automático, estos algoritmos suelen tener una tasa de precisión del 80% en la detección del discurso de odio.¹¹⁴ Esto implica que, aproximadamente 20% de los contenidos no se moderan adecuadamente; lo cual puede incluir contenido de odio que no fue detectado o contenido legítimo marcado como dañino. Tales fallas en la moderación automática de cientos de discursos legítimos generan ciertas preocupaciones en torno a la censura y la limitación de la expresión.¹¹⁵

¹⁰⁶ Larrondo & Garandi, *supra* nota 68, 4 del PDF.

¹⁰⁷ *Id.*

¹⁰⁸ *Id.*

¹⁰⁹ Durán Allimant, *supra* nota 70, en la págs. 7-8 del PDF.

¹¹⁰ Crosset & Dupont, *supra* nota 7, en la pág. 6 del PDF.

¹¹¹ *Id.*

¹¹² *Id.*

¹¹³ ONU, *supra* nota 15, en la pág. 17.

¹¹⁴ Sean MacAvaney, et al., *Hate speech detection: Challenges and solutions*, PLOS ONE, 20 de agosto de 2019, en la pág. 6 del PDF, <https://journals.plos.org/plusone/article/file?id=> (el estudio aborda los desafíos técnicos y las limitaciones de los algoritmos automatizados en la detección del discurso de odio en línea.).

¹¹⁵ DL Cade, *Instagram Says Its Anti-Spam System 'Incorrectly' Blocked Black Lives Matter Posts*, PETA PIXEL (2 de junio de 2020), <https://petapixel.com/2020/06/02/instagram-says-its-anti-spam-system-incorrectly-blocked-black-lives-matter-posts/>.

Uno de los mayores desafíos es la variabilidad lingüística. Aunque de cierta forma la IA tiene la habilidad de ejecutar tareas e imitar capacidades humanas,¹¹⁶ los algoritmos carecen de *corpus* y *mente*.¹¹⁷ Es decir, la IA aún no es capaz de comprender cuándo una expresión es irónica, una parodia o el contenido en distintos idiomas donde puede fallar al entender si un mensaje es inofensivo o dañino.¹¹⁸ Por lo tanto, las herramientas de automatización en su operatividad matemática tiende más a optar por un resultado rápido que consiste en limitar o remover la determinada expresión en línea sin tener en cuenta que ello afecta considerablemente al derecho humano a recibir, investigar y difundir información.¹¹⁹ Esto acarrea otras consecuencias, pues la censura excesiva y la minimización de la exposición a puntos de vista diversos, puede intensificar los sesgos e interferir con la iniciativa individual de cada usuario en la búsqueda de intercambiar ideas y opiniones entre diversos grupos ideológicos, políticos o sociales;¹²⁰ lo que se traduce en un impacto en la libertad de expresión de cada usuario.

Además, plantea ciertas preocupaciones en cuanto a la existencia de un equilibrio, o ausencia de éste, entre las decisiones humanas y las automatizadas, así como los problemas que surgen cuando las plataformas dependen exclusivamente de IA; pues no se sabe hasta qué punto las plataformas utilizan la automatización sin la intervención humana.¹²¹ Tal intervención algorítmica podría permitir a las plataformas sociales ejercer cierto grado de control sin precedentes sobre las comunicaciones públicas y privadas en Internet y, en su consecuencia, conducir a la censura automatizada.¹²²

III. IMPACTO DE LA IA EN LA LIBERTAD DE EXPRESIÓN: ALGUNOS POSIBLES PROBLEMAS JURÍDICOS RELACIONADOS A LA MODERACIÓN AUTOMATIZADA

A. Falta de transparencia:

La moderación de contenido en línea constituye un tema crucial que incide directamente sobre los derechos fundamentales y los valores democráticos de muchos usuarios. Esto se debe, en gran parte, a que las plataformas digitales establecen de forma autónoma estándares para la eliminación de contenido a nivel global.¹²³ Durante la última década, esta moderación de información digital se ha intensificado tanto en métodos como en la variedad de intervenciones editoriales y monitoreo en línea debido a las nuevas normas adoptadas por las plataformas las cuales, en algunos casos, responden a presiones estatales o demandas de la sociedad civil.¹²⁴

¹¹⁶ Pérez Martínez, *supra* nota 13, en la pág. 394.

¹¹⁷ Larrondo & Grandi, *Libertad y expresión*, *supra* nota 68, en la pág. 49 del PDF; ONU, *supra* nota 15, en la pág. 9.

¹¹⁸ Larrondo & Grandi, *Libertad y expresión*, *supra* nota 68, en la pág. 49 del PDF; Congressional Research Services, *supra* nota 3.

¹¹⁹ Larrondo & Grandi, *Libertad y expresión*, *supra* nota 68, en la pág. 49 del PDF.

¹²⁰ *Id.* en la pág. 48 del PDF.

¹²¹ Véase Schneider, *supra* nota 99; ONU, *supra* nota 15.

¹²² Véase Cobbe, *supra* nota 82.

¹²³ De Gregorio, *supra* nota 10, en la pág. 3 del PDF.

¹²⁴ Véase Martín Becerra & Silvio Waisbord, *La necesidad de repensar la ortodoxia de la libertad de expresión en la comunicación digital*, 60 DESARROLLO ECONÓMICO 232, 301 del PDF (2021), <https://www.jstor.org/stable/27032773>.

Sin embargo, estas nuevas reglas de edición, aplicadas de forma más estricta, no han sido acompañadas por mecanismos suficientes de protección para los usuarios.¹²⁵ Como resultado, numerosos contenidos han sido silenciados, ocultados o eliminados, y cuentas de personas u organizaciones han sido canceladas, sin garantías mínimas de apelación, transparencia o auditoría.¹²⁶ Esta opacidad en los sistemas de IA utilizados por las plataformas sociales ha sido señalada por diversos sectores dado a que tales métodos de moderación no permiten a los usuarios saber cómo y por qué se toman decisiones sobre el contenido que publican.¹²⁷

A pesar de que tienen un papel crucial en el flujo de información del ciberespacio, las plataformas de redes sociales no están obligadas a garantizar la transparencia ni ofrecer una explicación de sus procesos de toma de decisiones.¹²⁸ De manera que, y desde una perspectiva de la libertad de expresión del usuario a escala global, tal problema todavía no se ha resuelto.¹²⁹ Los informes de transparencia de las plataformas significaron un paso de la rendición de cuentas *ex post* de las políticas de moderación de contenido, sin embargo, tal y como señalan académicos y organizaciones, al final son insatisfactorios desde la perspectiva del usuario.¹³⁰

Por ejemplo, en el contexto de la Pandemia del Covid-19, el 10 de mayo de 2020, mediante un anuncio en su blog, sin más, Twitter actualizó su política de moderación de contenido en cuanto al contenido engañoso o de desinformación.¹³¹ En septiembre de 2022, Twitter anunció que suspendió más de 11,000 cuentas por violar las reglas de desinformación sobre el Covid-19 y eliminó 100,000 contenidos que infringían esta regla.¹³² Aunque para algunos fueron elogiadas tales estadísticas a raíz de las nuevas normas,¹³³ otros grupos mostraron su inconformidad con la opacidad del proceso a la luz de la libertad de expresión, pues en aquel momento las reglas no eran claras sobre cómo Twitter identificaría el contenido de “desinformación o engaño”, el “contenido polémico” o el contenido “no verificado” que categorizó en su comunicado.¹³⁴

Esta problemática se hizo más prominente cuando, en el año 2020, Facebook eliminó una publicación realizada por la página de Black Zebra Productions, una organización de periodismo independiente que produce documentales y transmisiones noticiosas en vivo ampliamente vistas en Internet.¹³⁵ En particular, Black Zebra publicó en su página de Face-

¹²⁵ *Id.* en la pág. 303.

¹²⁶ *Id.*

¹²⁷ Kai Riemer and Sandra Peter, *Algorithmic audiencing: Why we need to rethink free speech on social media* 36(4) *JIT* 409, 417 (2021), <https://journals.sagepub.com/doi/10.1177/02683962211013358>.

¹²⁸ Véase De Gregorio, *supra* nota 10.

¹²⁹ Becerra, *supra* nota 124, en la pág. 299 del PDF.

¹³⁰ *Id.* en las págs. 303-04 del PDF.

¹³¹ *Id.* en la pág. 304 del PDF.

¹³² Donie O'Sullivan, *Twitter ya no hará cumplir su política de desinformación sobre el covid-19, 29 de noviembre de 2022*, CNN (29 de noviembre de 2022), <https://cnnespanol.cnn.com/2022/11/29/twitter-ya-no-hara-cumplir-su-politica-de-desinformacion-sobre-el-covid-19-trax/>.

¹³³ *Id.*

¹³⁴ Becerra, *supra* nota 124, en la pág. 304 del PDF.

¹³⁵ *ACLU Demands Facebook Explain Why Content from Black Activists is Being Silenced*, ACLU NORTHERN CALIFORNIA (31 de julio de 2020), <https://www.aclunc.org/news/aclu-demands-facebook-explain-why-content-black-activists-being-silenced>.

book un vídeo que documentaba una protesta en Sacramento, el cual mostraba manifestaciones locales tras la muerte del Sr. George Floyd y de otras personas negras a manos de la policía de Minneapolis.¹³⁶ Dos días después, la página de esta organización fue desactivada por presuntamente violar los términos de la plataforma al publicar *contenido fraudulento o engañoso*, lo que llevó a Facebook a *moderar el contenido*.¹³⁷ Black Zebra Productions expresó que apeló la censura de su página, pero en ese momento no recibió respuesta de Facebook.¹³⁸ Posteriormente, la plataforma declaró que la página de la organización había sido marcada incorrectamente por supuestas violaciones de derecho de propiedad intelectual y que la publicación fue anulada por error.¹³⁹

Una situación similar ocurrió en Instagram, cuando los usuarios descubrieron que sus publicaciones con la etiqueta *#blacklivesmatter* eran bloqueadas.¹⁴⁰ Al intentar usar la etiqueta, recibían un mensaje emergente que decía: “[p]or favor, inténtelo de nuevo más tarde. Restringimos cierto contenido y acciones para proteger a nuestra comunidad”.¹⁴¹ Instagram reconoció el problema y explicó mediante un comunicado que su tecnología de detección de *spam* estaba operando incorrectamente debido al alto volumen de publicaciones con ese *hashtag*.¹⁴² Sin embargo, la empresa no aclaró si había resuelto el problema, así como tampoco reveló cuántos usuarios o publicaciones podrían haber resultado afectados.¹⁴³

B. Sesgos algorítmicos:

De entrada, debemos establecer que, aun cuando pudiera tener sus ventajas, la IA no es inherentemente imparcial. Al respecto, la investigadora norteamericana Safiya Umoja Noble, nos plantea que los algoritmos, lejos de tener un carácter neutral y objetivo, tienen una fuerte base en su funcionamiento que conducen a resultados sesgados, sexistas, racistas o de falsas nociones de meritocracia.¹⁴⁴ Ello, se hace más prominente cuando reconocemos que los algoritmos se han vuelto cada vez más normativos y son parte de las experiencias digitales y cotidianas de la población en la medida que sobre tal tecnología descansa la tarea de permitir o no el acceso a múltiples servicios o a información de diverso carácter.¹⁴⁵ El punto de partida de esta problemática es el conjunto de instrucciones o reglas matemáticas impregnadas en sus diseños. Los algoritmos que impulsan la IA dependen de datos de en-

¹³⁶ Michael McGough, *Black Zebra, team covering Sacramento protests, after Facebook disabled it*, THE SACRAMENTO BEE (9 de junio de 2020), <https://amp.sacbee.com/news/local/article243401291.htm>.

¹³⁷ *Id.*

¹³⁸ Pema Levy, *Black Activists Warn that Facebook Hasn't Done Enough to Stop Racist Harassment*, MOTHER JONES (9 de julio de 2020), <https://motherjones.com/politics/2020/07/facebook-black-lives-matter>.

¹³⁹ *Id.*

¹⁴⁰ DL Cade, *supra*, nota 115.

¹⁴¹ *Id.* (traducción suplida).

¹⁴² *Id.*

¹⁴³ *Id.*

¹⁴⁴ Juan Camilo Gómez Barrera, *Segmentación, sesgo y normas sociales en la programación: aportes a la teoría de la gubernamentalidad algorítmica*, 15 AVATARES DE LA COMUNICACIÓN Y LA CULTURA 111 DEL PDF (2018), <https://publicaciones.sociales.uba.ar/index.php/avatares/article/view/4889/4022>.

¹⁴⁵ *Id.*

trenamiento insertados por humanos utilizados para crear sus modelos.¹⁴⁶ De este modo, si estos datos reflejan sesgo sociales o discriminatorios preexistentes, es probable que el algoritmo los reproduzca o, incluso, los amplifique.¹⁴⁷ Tomemos como ejemplo a Amazon, su sistema de reclutamiento de IA fue entrenado con contrataciones realizadas durante los últimos diez años, en las cuales eran predominante los hombres.¹⁴⁸ Como resultado, el algoritmo aprendió que los hombres eran mejores candidatos y descartó automáticamente los currículos de candidatos con nombres o características femeninas.¹⁴⁹

Asimismo, recientemente la UNESCO realizó una investigación sobre los estereotipos de género, origen racial o identidad sexual y concluyó que los sistemas de IA como GPT-2 y Llama asociaban desproporcionadamente los nombres de mujer con palabras como hogar, familia, hijos o matrimonio.¹⁵⁰ En cuanto a los nombres de hombres, la IA los vinculó con significados como ejecutivo, salario y carrera.¹⁵¹ Cuando se les pidió a los sistemas que completaran una oración que mencionara el género de una persona junto con su identidad sexual, aproximadamente el 20% incluía lenguaje sexista o misógino y, en el 60% de los casos, generó contenido negativo sobre las personas homosexuales.¹⁵² No hay duda de que los sesgos son un desafío significativo para los algoritmos de censura en general.¹⁵³ A pesar de ello, en la medida en que estos sesgos están impregnados en las variables utilizadas por la IA, esta podrá acabar derivando y acentuando sesgos sistemáticos.¹⁵⁴

Estos sesgos generados por los algoritmos tienen consecuencias fatales en la sociedad y la presunción de objetividad que viene con las herramientas tecnológicas puede hacer que las personas estén menos dispuestas a reconocer tal problema.¹⁵⁵ Se ha demostrado cómo estas herramientas tecnológicas pueden perpetuar los prejuicios relacionados con el género, la raza, la religión, la afiliación política y afectar desproporcionadamente a grupos ya marginados.¹⁵⁶ Particularmente, al analizarse un conjunto de datos y clasificadores populares de discurso de odio en la red social Twitter, se encontró que los tuits de usuarios afroamericanos tenían hasta dos veces más probabilidades de ser etiquetados como *ofen-*

¹⁴⁶ *When AI gets it wrong: addressing AI hallucinations and bias*, MIT MANAGEMENT STS TEACHING & LEARNING TECHNOLOGIES, <https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/> (última visita 19 de diciembre de 2024).

¹⁴⁷ Carolina Aliaga, *Inteligencia artificial y sesgos de género: Entrenar a los algoritmos para no perpetuar discriminaciones*, UNIVERSIDAD DE CHILE (12 de marzo de 2024), <https://uchile.cl/noticias/213982/avances-y-desafios-sobre-inteligencia-artificial-y-sesgos-de-genero>.

¹⁴⁸ Jeffrey Dastin, *Insight - Amazon scraps secret AI recruiting tool that showed bias against women*, REUTERS (10 de octubre de 2018), <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCNiMKoAG/>.

¹⁴⁹ *Id.*

¹⁵⁰ UNESCO, CHALLENGING SYSTEMATIC PREJUDICES: AN INVESTIGATION INTO BIAS AGAINST WOMEN AND GIRLS IN LARGE LANGUAGE MODELS 9 del PDF (2024), <https://unesdoc.unesco.org/ark:/48223/pf0000388971>.

¹⁵¹ *Id.*

¹⁵² *Id.* en la pág. 10.

¹⁵³ Cobbe, *supra* nota 82, en la pág. 754 del PDF.

¹⁵⁴ Ana Ginés i Fabrellas, *Los algoritmos están impulsando la desigualdad, no eliminándola*, DO BETTER (5 de septiembre de 2024), <https://dobetter.esade.edu/es/algoritmos-desigualdad#:~:text=En%20lugar%20de%20eliminar%20las%20desigualdades%20al%20entregar,orientaci%C3%B3n%20sexual%20y%20%20a%20un%20ritmo%20alarmante>.

¹⁵⁵ MIT MANAGEMENT, *supra* nota 146.

¹⁵⁶ Véase *Id.*

sivos que los tuits de otros usuarios.¹⁵⁷ Esto no solo obstaculiza el potencial de las redes sociales para dar voz a las comunidades marginadas, sino que tal represión podría exacerbar los sentimientos de discriminación de estos grupos.¹⁵⁸

A su vez, los algoritmos de moderación en plataformas digitales carecen de los matices del contexto humano.¹⁵⁹ Es decir, la IA puede estar sesgada por el déficit de información contextual ya que cuenta, al presente, con limitaciones importantes, entre las que se destaca la imposibilidad de poder evaluar el marco de referencia de las expresiones, “los usos idiomáticos y [los] aspectos culturales de los seres humanos”.¹⁶⁰ En muchas ocasiones, los algoritmos no son capaces de identificar la sátira, el contexto cultural, la intención del usuario, el tono de la publicación, al público que va dirigido el contenido o el propósito artístico del contenido a moderar.¹⁶¹ Como resultado, existe el riesgo de una interpretación y clasificación errónea, lo que, como consecuencia, podría llevar a la censura de expresiones locales o coloquiales, particularmente en los casos dónde el contexto es crucial para determinar su permisibilidad.¹⁶²

Por ejemplo, el Sr. Alexis Zárraga Vélez, escritor, columnista y creador del podcast *Siempre es lunes*, explicó cómo fue censurado en redes sociales por utilizar palabras como *machorra*, la cual comúnmente menciona en algunos de sus podcasts, por estos sistemas operar sin contexto cultural.¹⁶³ A pesar de que en Puerto Rico este término no necesariamente tiene una connotación negativa y puede referirse a algo distinto, el algoritmo lo interpretó como una expresión ofensiva.¹⁶⁴ Lo mismo ocurre con palabras como *negrito*, que en ciertos contextos es una expresión de cariño, pero los sistemas automatizados de plataformas como Facebook e Instagram las marcan como inapropiadas o peyorativas, ig-

¹⁵⁷ Cobbe, *supra* nota 82, en las págs. 754-55 del PDF (traducción suplida).

¹⁵⁸ Véase Cino Lee, et al., *People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise*, 121 PNAS 1 (2024), <https://www.pnas.org/doi/full/10.1073/pnas.2322764121>.

¹⁵⁹ ONU, *supra* nota 15, en las págs. 9, 13 del PDF.

¹⁶⁰ Manuel Ernesto Larrondo & Nicolás Mario Grandi, *Inteligencia Artificial, algoritmos y libertad de expresión*, 34 *Universitas* 177, 183 del PDF (2021), http://scielo.senescyt.gob.ec/scielo.php?script=sci_arttext&pid=S1390-86342021000100177; [en adelante, “Larrondo & Grandi, *Inteligencia Artificial*”]; Véase también Nicholas Ruiz & Marcello Federico, *Assessing the Impact of Speech Recognition Errors on Machine Translation Quality*, 1 ACL ANTHOLOGY 261, 262 del PDF (2014), <https://aclanthology.org/2014.amta-researchers.20/>. (El estudio nos ilustra sobre el impacto de los errores de reconocimiento de voz (ASR) en la traducción automática (MT). Los autores señalaron que aproximadamente el 50% de los errores que comete esta tecnología están asociados a la sustitución incorrecta de palabras, a menudo relacionadas con la morfología de las palabras.).

¹⁶¹ XIANHONG HU ET AL., EL APOORTE DE LA INTELIGENCIA ARTIFICIAL Y LAS TIC AVANZADAS A LAS SOCIEDADES DEL CONOCIMIENTO: UNA PERSPECTIVA DE DERECHOS, APERTURA, ACCESO Y MÚLTIPLES ACTORES 45 del PDF (2021), <https://unesdoc.unesco.org/ark:/48223/pf0000375796>; Véase también Juan Londoño, *Assessing the Impact of the Widespread Adoption of Algorithm-backed Content Moderation in Social Media*, AMERICAN ACTION FORUM (25 de enero de 2022), <https://www.americanactionforum.org/insight/assessing-the-impact-of-the-widespread-adoption-of-algorithm-backed-content-moderation-in-social-media/>.

¹⁶² Véase Minas Stravopodis, *AI & Freedom of Expression in the Contemporary Digital Landscape*, IRIS SUSTAINABLE DEVELOPMENT (2024), https://www.irissd.org/post/ai-freedom-of-expression-in-the-contemporary-digital-landscape?utm_source=chatgpt.com.

¹⁶³ Entrevista con el Sr. Alexis Zárraga Vélez a través de WhatsApp (en adelante, “señor Zárraga Vélez”), escritor, columnista y creador de contenido en podcast (28 de octubre de 2024).

¹⁶⁴ *Id.*

norando así la rica variabilidad cultural en el lenguaje.¹⁶⁵ Debido a este sesgo en el tratamiento del lenguaje, muchos terminan adaptando su forma de expresarse para mejorar su interacción en la tecnología.¹⁶⁶

C. La autocensura y el debate público:

Al comienzo de su desarrollo, los fundadores y partidarios de las redes aspiraban a que las plataformas, como Facebook y Twitter, produjeran espacios para el debate de aquellos que tenían opiniones minoritarias, de manera que se sintieran libres de expresar sus ideas, ampliando así el discurso público y añadiendo nuevas perspectivas al debate cotidiano sobre cuestiones políticas.¹⁶⁷ Sin embargo, la creciente implementación de la moderación automatizada ha introducido una nueva capa de complejidad en la libertad de expresión y el debate público: la autocensura.

La autocensura puede definirse como el acto de limitar o controlar intencional y voluntariamente la propia expresión, no debido a las restricciones impuestas directamente por la plataforma, sino con el propósito de evitar ofender o incomodar a otros, o para ajustarse a las normas sociales o culturales.¹⁶⁸ Ello, ha traído consigo desafíos complejos, entre estos se encuentra la censura preventiva y la amplificación algorítmica de la percepción de la mayoría, una especie de combinación entre la teoría de la *espiral del silencio* y las *burbujas de filtro*. Ambos fenómenos, explicados a continuación, inciden directamente en la autocensura y configuran un entorno en el que las herramientas de IA para moderación de contenido, aunque útiles,¹⁶⁹ puedan influir involuntariamente en el discurso público, limitando la libertad de expresión y reduciendo el pluralismo que las plataformas originalmente pretendían fomentar.¹⁷⁰

165 *Id.* (En cuanto a la sátira, el señor Zárraga Vélez afirmó que el algoritmo “no entiende de contexto, del [argot] de un país, ni comprende lo denso que es el idioma español”. Por lo que, entiende que al algoritmo “habría que ‘humanizarlo’, aunque [aclaró que] a veces la sátira no la entienden ni los humanos”).

166 *Id.* (Durante la entrevista, el señor Zárraga Vélez explicó que, aunque no practica la autocensura, reconoce que la falta de precisión en el lenguaje por parte de los algoritmos ha moldeado su enfoque al publicar ciertos contenidos. En particular, comentó que a veces es necesario ajustar un poco las palabras, usando sinónimos o expresiones alternativas, para transmitir lo mismo y así superar el algoritmo.). Véase también Zion Mengesha et al., “I don’t think these devices are very culturally sensitive.”-Impact of Automated Speech Recognition Errors on African Americans, 4 FRONTIERS IN ARTIFICIAL INTELLIGENCE 1, 9 DEL PDF (2021), <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.725911/ful>. (El referido estudio contempla las dificultades que enfrentan los hablantes afroamericanos al utilizar sistemas de reconocimiento de voz (ASR) y señala la necesidad de diseñar sistemas que sean más justos, inclusivos y equitativos para todos los usuarios, independientemente de su origen étnico o dialecto.).

167 KEITH HAMPTON ET AL., SOCIAL MEDIA AND THE ‘SPIRAL OF SILENCE’, 3 del PDF (2014), <https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/>.

168 Véase, Wei Hu & Diogo Barradas, *Work in progress: a glance at social media self-censorship in North America*, 2023 IEEE EUROPEAN SYMPOSIUM ON SECURITY AND PRIVACY WORKSHOPS (EuroS&PW) (31 de julio de 2023), <https://ieeexplore.ieee.org/document/10190716>.

169 CONGRESSIONAL RESEARCH SERVICES, *supra* nota 3.

170 HAMPTON ET AL, *supra* nota 167.

i. Amplificación algorítmica: percepción de la mayoría y las burbujas de filtro

La autocensura, alimentada por la desaprobación social es en un mecanismo de autorregulación que mantiene activa la *espiral del silencio* (*spiral of silence*). En el 1974, la socióloga Noelle-Neumann, en su artículo *The Spiral of Silence a Theory of Public Opinion*, elaboró la teoría de la *espiral del silencio*; la cual explica cómo la percepción de una mayoría dominante y el miedo al aislamiento pueden llevar a la autocensura.¹⁷¹ Su análisis sugiere que las personas tienden a callar sus opiniones si perciben que son minoritarias o socialmente impopulares.

Dentro del contexto de las redes sociales y la moderación automatizada, diversos estudios han confirmado que los efectos del *espiral del silencio* están presentes en el internet.¹⁷² De acuerdo con un informe publicado por Pew Research Center, titulado *Social Media and the 'Spiral of Silence'*, los efectos de la espiral de silencio observados anteriormente en la disposición de las personas para expresarse en diversos entornos fuera de línea también se manifiestan entre los usuarios de plataformas sociales.¹⁷³ Ello generó diversas discusiones al poner en evidencia la preocupante realidad de que las redes sociales están fomentando la tendencia de los usuarios a autocensurarse y evitar expresar opiniones cuando perciben que son minoritarias.¹⁷⁴ -

En este contexto, el fenómeno de la *espiral del silencio*, que combina la percepción de la opinión mayoritaria con la autocensura, se intensifica en las redes sociales mediante el efecto de las *burbujas de filtro*.¹⁷⁵ Las *burbujas de filtro* son algoritmos diseñados para mostrar contenido que refuerza las creencias del usuario y reducen la exposición de ideas contrarias.¹⁷⁶ Las redes sociales utilizan métodos de análisis inferencial para anticipar las preferencias de los usuarios, lo que puede incluir atributos como raza, género, orientación sexual y posturas políticas.¹⁷⁷ Estos algoritmos, aunque ricos en información y diseñados para personalizar la experiencia,¹⁷⁸ pueden generar burbujas de información capaces de reforzar sesgos, amplificar la percepción de la mayoría y limitar la exposición a diferentes perspectivas.¹⁷⁹ Sandra

¹⁷¹ Elisabeth Noelle-Neumann, *The spiral of silence a theory of public opinion*, 24 JOURNAL OF COMMUNICATION 43 (1974); <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1974.tb00367.x> (1974); Véase también ALEXIS Figueroba, *Espiral del silencio: ¿qué es y cuáles son sus causas?*, PSICOLOGÍA Y MENTE (24 de junio de 2017), <https://psicologiaymente.com/social/espiral-del-silencio>.

¹⁷² Anna Gabibson, et al., *Free speech and safe spaces: how moderation policies shape online discussion spaces*, 5 SOCIAL MEDIA + SOCIETY 1, 3 (2019), <https://journals.sagepub.com/doi/10.1177/2056305119832588#bibr62-2056305119832588>.

¹⁷³ HAMPTON ET AL, *supra* nota 167, en las págs. 3-4 del PDF.

¹⁷⁴ Véase Enrique Dans, *Redes Sociales y autocensura*, ED (28 de agosto de 2014), <https://www.enriquedans.com/2014/08/redes-sociales-y-autocensura.html>.

¹⁷⁵ Eduardo Fosch-Villaronga et al., *Gendering algorithms in social media*, ACM SIGKDD EXPLORATIONS NEWSLETTER 24, 25 (2021), <https://dl.acm.org/doi/10.1145/3468507.3468512>.

¹⁷⁶ Aníbal Rossi, *¿Burbujas de filtro? Hacia una fenomenología algorítmica*, 13 INMEDIACIONES de la Comunicación 256, 264-65 (2018), <https://www.redalyc.org/pdf/5897/589776630012.pdf>. (De acuerdo con el escritor estadounidense Eli Pariser, el fenómeno de "burbujas de filtros" se refiere "al aislamiento informativo donde estarían quedando atrapados los usuarios como consecuencia de dichas configuraciones").

¹⁷⁷ Fosch-Villaronga et al., *supra* nota 175.

¹⁷⁸ Londoño, *supra* nota 161.

¹⁷⁹ *La verdad tras las burbujas de filtro: desmontando algunos mitos*, FUNDACIÓN LUCA DE TENA (28 de enero de 2020), <https://laboratoriodeperiodismo.org/desmontando-mitos-sobre-las-burbujas-de-filtro/>.

Álvaro nos ilustra con el ejemplo de Facebook, quien tiene un algoritmo llamado *Edgerank* que analiza nuestros datos de navegación como los *me gusta* que ponemos, los amigos que tenemos y los comentarios que hacemos.¹⁸⁰ Con dicha información Facebook nos perfila con el fin de mostrarnos aquellas historias que nos gustan y ocultarnos aquello que nos aburre y mostrarnos nuevos amigos que concuerdan con nuestro perfil e ideología.¹⁸¹

Ciertamente, la personalización del contenido en línea por parte de la IA ofrece información y conexiones relevantes. Sin embargo, estudios demuestran que mediante este tipo de práctica se acentúan los sesgos de confirmación, ya que las recomendaciones algorítmicas provienen de fuentes menos diversas y pueden reforzar los sesgos de percepción y la falsa idea de estar en minoría.¹⁸² La UNESCO advierte que tal práctica “puede manipular la forma en que las personas usan su derecho a buscar información y su derecho a formarse una opinión”.¹⁸³ Como resultado, podría disminuir el pluralismo de pensamientos y el grado de exposición a la información válida.¹⁸⁴

Siguiendo la línea de pensamiento de Eli Pariser, el requisito de la democracia deliberativa no es sólo que los ciudadanos tengan libertad de expresión y el derecho al pluralismo de ideas, sino que estén efectivamente expuestos a las opiniones y a diversos puntos de vista.¹⁸⁵ Estar cerca de las *burbujas de filtros* no permite esta exposición en un grado significativo. Por lo tanto, es posible que, aunque las redes sociales no bloquean directamente el acceso a cierta información, sino que, de alguna forma, lo oculten a los usuarios, esto podría ser suficiente para representar una amenaza para el funcionamiento efectivo del intercambio de opiniones y pensamientos y, por lo tanto, para la libertad de expresión y la democracia deliberativa en general.¹⁸⁶

ii. Censura preventiva

La autocensura preventiva en la moderación del contenido en línea es un fenómeno creciente, influenciado por diversas políticas de moderación y el temor de los usuarios a repercusiones sociales o legales.¹⁸⁷ Esto puede ocurrir generalmente en temas controverbiales o sensibles que, aunque no necesariamente infringen las normas de la plataforma, corren el riesgo de ser mal interpretados por los sistemas automatizados que operan bajo parámetros rígidos u opacos. Las plataformas sociales, a través de sus directrices comunitarias, pueden influir y hasta cierto punto orientar a los usuarios sobre qué pueden decir,

¹⁸⁰ Álvaro, *supra* nota 75.

¹⁸¹ *Id.*

¹⁸² Andrés Abeliuk, *Sesgos Algorítmicos en las Redes Sociales*, 24 REV. BITS DE CIENCIA 2, 3-4 del PDF (2013), 12669-Texto del artículo-30608-1-10-20230620.pdf.

¹⁸³ XIANHONG HU ET AL., *supra* nota 161, en la pág. 19 del PDF.

¹⁸⁴ *Id.*

¹⁸⁵ Krzysztof Jankowski, *Living in the Filter Bubble: Is What We Lose Something We Need to Preserve?* SSRN 1 del PDF (24 de noviembre de 2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2982025. (citando a ELI PARISER, *THE FILTER BUBBLE. WHAT THE INTERNET IS HIDING FROM YOU* (2017)).

¹⁸⁶ *Id.*

¹⁸⁷ Elvin Ong, *Online repression and self-censorship: evidence from Southeast Asia*, 56 GOVERNMENT AND OPPOSITION 141, 142 (2021), <https://www.cambridge.org/core/journals/government-and-opposition/article/abs/online-repression-and-self-censorship-evidence-from-southeast-asia/5ADD8CE94376F3E4707A1D833EC6180C>.

compartir o, incluso, aquello que no deben expresar en las redes sociales.¹⁸⁸ Debido a la naturaleza opaca, la inestabilidad y la agregación discrecional de criterios sobre el contenido sancionado con remoción o invisibilidad por parte de las plataformas, los usuarios tienden a autocensurarse para evitar represalias, tratando de intuir qué contenido sería ‘aceptable’ conforme a las cambiantes reglas de las redes.¹⁸⁹ Aun cuando la autocensura es un elemento difícil de evaluar, es probable que las normas que rigen las redes sociales impongan una limitación en el discurso que se desarrolla en ellas, lo que puede hacer que los usuarios sientan que ciertos temas que no deben ser abordados.¹⁹⁰

Este fenómeno se ha vuelto un elemento central en la configuración del debate público.¹⁹¹ En su artículo, *Self-Censorship: The Chilling Effect and the Heating Effect*, Robert Mark Simpson, analiza cómo las restricciones legales al discurso público pueden generar, entre otras cosas, un efecto disuasorio (*chilling effect*), donde el temor a sanciones lleva a los individuos a autocensurarse, incluso más allá de lo que la ley prohíbe.¹⁹² Aunque el autor se enfoca en las restricciones impuestas por el Estado, su análisis es particularmente relevante en el contexto de la moderación de contenido en plataformas digitales, dónde las políticas internas y los sistemas automatizados pueden inducir a los usuarios a restringir su propia expresión por temor a represalias.

El efecto disuasorio ocurre cuando los riesgos asociados a una restricción de la libertad de expresión, de forma involuntaria, desalienta el discurso que queda fuera del alcance oficial de dicha restricción.¹⁹³ En el caso de la moderación algorítmica, la censura automatizada de ciertos contenidos en internet podría desalentar el uso de las redes sociales como un canal para difundir ideas, críticas e información. En ese sentido, algunos usuarios comparten como, debido a bloqueos excesivos por parte de ciertas plataformas, optan por cerrar sus cuentas para evitar que sus publicaciones sean eliminadas por las normas y políticas de moderación.¹⁹⁴ En consecuencia, el espacio de expresión en las redes sociales se transforma en un entorno donde los mismos usuarios filtren sus opiniones,¹⁹⁵ limitando las discusiones auténticas y reduciendo la diversidad de perspectivas.

En fin, los fenómenos de la autocensura en combinación a la amplificación algorítmica, a través de la *espiral del silencio* y las *burbujas de filtro*, revelan cómo el temor a la restricción del discurso y los sesgos algorítmicos pueden limitar la diversidad de opiniones

¹⁸⁸ TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 71 del PDF (2018), file:///C:/Users/ngd4o6o/Downloads/Gillespie_CUSTODIANS_print.pdf.

¹⁸⁹ Riemer & Peter, *supra* nota 127, en la pág. 422 del PDF.

¹⁹⁰ GILLESPIE, *supra* nota 188.

¹⁹¹ *Id.*

¹⁹² Robert Mark Simpson, *Self-Censorship: The Chilling Effect and the Heating Effect*, 1(2) POLITICAL PHILOSOPHY 345, 346-47 del PDF (2024), <https://philarchive.org/rec/SIMSTC-2>. (Aunque el autor analiza en su escrito el efecto disuasorio (*chilling effect*) en el contexto de las restricciones estatales al discurso, su teoría puede aplicarse de forma ilustrativa a las políticas y normas de moderación de contenido en las redes sociales, dónde la incertidumbre, falta de claridad y transparencia pueden inducir a la autocensura entre los usuarios.).

¹⁹³ *Id.*

¹⁹⁴ Véase Waller, *supra* nota 65.

¹⁹⁵ Maialen Ferreira, *El terror sexual, la autocensura y el odio en redes, a debate en el III Congreso de Periodismo Feminista*, EL DIARIO (16 de octubre de 2023), https://www.eldiario.es/euskadi/terror-sexual-autocensura-odio-redes-debate-iii-congreso-periodismo-feminista_1_10602418.html.

y crear entornos de pensamiento uniforme, reduciendo el acceso a perspectivas variadas. Como resultado de tales prácticas, se corre el riesgo de bloquear la legítima libertad de expresión y afectar negativamente la diversidad de opiniones en el debate público.¹⁹⁶

IV. LA VISIÓN INTERNACIONAL: PRINCIPIOS DE DERECHOS HUMANOS PARA LA MODERACIÓN DE CONTENIDO EN LAS REDES SOCIALES

La libertad de expresión es un derecho que no tiene fronteras. Así lo reconoce tanto el artículo 19(2) del Pacto Internacional de Derechos Civiles y Políticos, como el artículo 13(1) de la Convención Americana de Derechos Humanos.¹⁹⁷ Ambas disposiciones rodean de amplias garantías el derecho a la libertad de expresión y establecen que este comprende “la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística, o por cualquier otro procedimiento a elección”.¹⁹⁸ Es decir, la libertad de expresión es inequívocamente un derecho humano básico que “no solo tiende a la realización personal de quien se expresa[,] sino a la consolidación de sociedades democráticas”.¹⁹⁹ Tal principio fundamental aplica tanto fuera de línea como en Internet.²⁰⁰

Ante ello, los Estados tienen la obligación de generar las condiciones necesarias “para una deliberación pública, plural y abierta sobre los asuntos que conciernen a todas y todos como ciudadanos”.²⁰¹ En cuanto a las empresas, el derecho internacional reconoce ciertos Principios Rectores que, si bien no son vinculantes, establecen “normas de conducta mundial aplicables a todas las empresas”, los cuales deben regir todas las operaciones, cualquiera que sea el lugar en el que operen.²⁰² Para ilustrar, ciertos Principios Rectores recomiendan a las empresas que adopten compromisos de política de alto nivel orientados a respetar los derechos humanos y que participen en estrategias de prevención y mitigación que respeten los derechos humanos.²⁰³

En los últimos años, ha tomado fuerza la idea de que las plataformas sociales deberían aplicar a sus reglas de moderación de contenido los principios del Derecho Internacional

¹⁹⁶ James L. Gabson & Joseph L. Sutherland, *Keeping your mouth shut: spiraling self-censorship in the United States*, 138 POLITICAL SCIENCE QUARTERLY 361, 362 del PDF (2023), <https://academic.oup.com/psq/article/138/3/361/71928889>.

¹⁹⁷ Paula Roko, *El control estatal de la (Des)información en Internet en el contexto de la pandemia: un análisis de las tendencias regionales bajo una perspectiva de libertad de expresión*, 37 AM. U. INT'L L. REV. 253, 285 (2022).

¹⁹⁸ A.G. Res. 2200 A (XXI), Pacto Internacional de Derechos Civiles, art. 19(2), (23 de marzo de 1976), <https://www.ohchr.org/es/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>; Convención Americana de Derechos Humanos (Pacto San José), OEA, art. 13(1), 22 de noviembre de 1969, ONU 27/08/1979 17955, https://www.oas.org/dil/esp/1969_Convenci%C3%B3n_Americana_sobre_Derechos_Humanos.pdf.

¹⁹⁹ Roko, *supra* nota 197, en la pág. 288.

²⁰⁰ *Id.* en las págs. 286-87.

²⁰¹ *Id.* en las págs. 288-89.

²⁰² Consejo de Derechos Humanos, Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y expresión, U. N. Doc. A/ HRC/38/35, en la pág. 6 del PDF (6 de abril de 2018), <https://www.refworld.org/es/ref/infortem/ccpr/2018/es/129062>.

²⁰³ *Id.*

sobre derechos humanos.²⁰⁴ Esta postura no solo fue planteada por una gran parte de la academia y la sociedad en general, sino que también fue promovida por el Informe del entonces Relator Especial de las Naciones Unidas en el 2018.²⁰⁵ Mediante dicho informe se advirtió sobre las medidas agresivas que tomaban las compañías para hacer frente a asuntos como la desinformación y la importancia de la promoción y protección del derecho de la libertad de expresión y pensamiento.²⁰⁶ El informe señala a las empresas de redes sociales como “reguladores enigmáticos, que establecen una especie de ‘ley de las plataformas’ en la que es difícil percibir elementos como claridad, coherencia, rendición de cuentas y reparación”.²⁰⁷

Es por lo cual, la ONU propone la adopción de un enfoque en los derechos humanos para establecer una base universal y de principios que garanticen que las medidas relacionadas con el contenido publicado por los usuarios se guíe por las mismas normas de legalidad, necesidad y legitimidad que rigen la regulación de la expresión por los Estados.²⁰⁸ Por ejemplo, para adaptar las conclusiones de los Principios Rectores al ámbito de la IA, la ONU recomienda que las empresas de redes sociales se comprometan a respetar los derechos humanos de los usuarios en las aplicaciones de la IA a través de políticas de alto nivel; eviten, mitiguen y prevengan cualquier impacto negativo en los derechos humanos derivado de la IA, y ofrezcan mecanismos accesibles de reclamación para remediar los efectos negativos en los derechos humanos generados por los sistemas de IA.²⁰⁹ Tales principios del derecho internacional se fundan en la idea de que la libertad de expresión es “una piedra angular en la existencia misma de una sociedad democrática”.²¹⁰

CONCLUSIÓN

La IA ha transformado radicalmente el panorama de la libertad de expresión, presentando tanto oportunidades como desafíos sin precedentes. Por un lado, ha contribuido a democratizar el acceso a la información ampliando las posibilidades de expresión, especialmente a través de las redes sociales, las cuales se han consolidado como uno de los principales canales de difusión de ideas e información.²¹¹ Por otro lado, estas mismas plataformas desempeñan un rol central en la lucha contra la difusión de material dañino. Ene se mismo marco, la moderación de contenido en línea no solo se ha vuelto necesaria, sino en ocasiones ineludible, con el fin de garantizar que los espacios digitales no se conviertan

²⁰⁴ Roko, *supra* nota 197, en la pág. 299; Véase también *Moderating online content: fighting harm or silencing dissent?*, United Nation Human Rights Office of the High Commissioner (23 de julio de 2021), <https://www.hayderecho.com/2022/05/20/el-impacto-de-la-inteligencia-artificial-en-la-libertad-de-expresion-y-de-pensamiento-el-papel-de-los-neuroderechos/>.

²⁰⁵ Roko, *supra* nota 197, en la pág. 299.

²⁰⁶ *Id.* en la pág. 301.

²⁰⁷ Consejo de Derechos Humanos, *supra* nota 202, en la pág. 3 del PDF.

²⁰⁸ *Id.* en las págs. 18.

²⁰⁹ ONU, *supra* nota 15, en las págs. 10-11.

²¹⁰ Roko, *supra* nota 197, en la pág. 288.

²¹¹ María Jesús González-Espejo, *El impacto de la Inteligencia Artificial en la libertad de expresión y de pensamiento. El papel de los neuroderechos*, HAY DERECHO (20 de mayo de 2022), <https://www.hayderecho.com/2022/05/20/el-impacto-de-la-inteligencia-artificial-en-la-libertad-de-expresion-y-de-pensamiento-el-papel-de-los-neuroderechos/>.

en entornos hostiles para los usuarios por la proliferación de conductas como la incitación al odio, la discriminación, incitación a la violencia o el plagio.

Ahora bien, si bien la moderación automatizada en redes sociales aspira a proteger a los usuarios, plantea un dilema ético fundamental entre la garantía de espacios digitales seguros y la preservación del derecho a la libertad de expresión.²¹² Aunque algunos promueven las herramientas automatizadas como la fuerza que, de alguna forma, salvará la moderación de sus propias limitaciones,²¹³ diversos estudios han puesto en evidencia los desafíos a los cuales se enfrenta la sociedad y el impacto que tiene en el ejercicio libre de este derecho fundamental.²¹⁴ De manera particular, la moderación automatizada puede conllevar la eliminación de opiniones disidentes válidas, promover la autocensura derivada de normas imprecisas, consolidar *burbujas de filtro* que limitan la exposición a ideas diversas y reproducir, e incluso amplificar, los sesgos impregnados en estos sistemas.

En consecuencia, las medidas dirigidas a mitigar los efectos del discurso considerado como perjudicial -como la eliminación de expresiones o el bloqueo de cuentas de usuarios, especialmente a gran escala-, pueden desafiar el derecho fundamental de todo ser humano a recibir e impartir información e ideas por cualquier medio.²¹⁵ Para que la IA contribuya efectivamente a la promoción de la democracia y los derechos humanos, resulta indispensable el desarrollo de marcos regulatorios sólidos, la promoción de la alfabetización digital, el fortalecimiento de las instituciones democráticas y el diseño de sistemas de moderación de contenido más transparentes, justos y sometidos a mecanismos adecuados de rendición de cuentas.

La libertad de expresión constituye una de las piedras angulares sobre la que se basa la democracia.²¹⁶ En tal virtud, solo mediante un enfoque equilibrado y responsable podrá asegurarse que las tecnologías emergentes, lejos de silenciar voces legítimas, sirvan para empoderar y promover el ejercicio de este derecho en el entorno digital.

²¹² De Gregorio, *supra* nota 10, en la pág. 4 del PDF.

²¹³ Véase Gorwa et al., *supra* nota 71.

²¹⁴ Véase ONU, *supra* nota 15, en las págs. 13-16 del PDF; UNESCO, *supra* nota 9; Larrondo & Grandi, *Libertad y Expresión*, *supra* nota 68; Álvaro, *supra* nota 75; Gorwa et al., *supra* nota 71.

²¹⁵ Kozyreva et al., *supra* nota 99.

²¹⁶ Claudio Grossman, *La libertad de expresión en el sistema interamericano de protección de los derechos humanos*, 7 ILSA J. INT'L & COMP. L. 159 DEL PDF (2001), <https://nsuworks.nova.edu/ilsajournal/vol7/iss3/11/>.