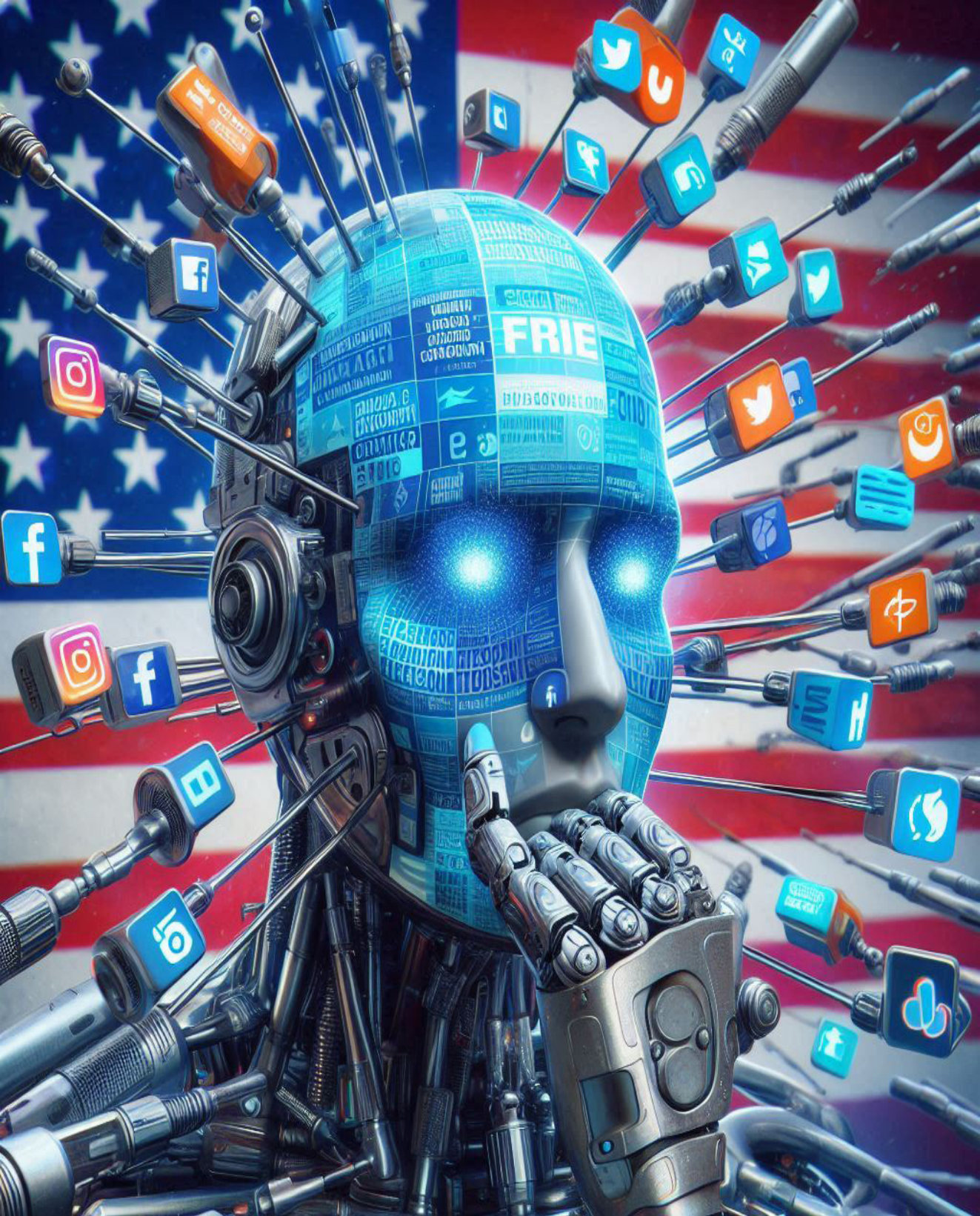
FRIE

LIBERTAD DE EXPRESIÓN EN LAS REDES SOCIALES: *ANÁLISIS DE LA CENSURA AUTOMATIZADA POR INTELIGENCIA ARTIFICIAL*

Mauricio José C. Rosales[1]
Gerardo José Gómez Ramírez[2]
Marcia Nallely Zúniga Velásquez[3]
Richard Said Salinas Aguiriano[4]
Lorena Argentina Sánchez Maradiaga[5]

DOI: https://doi.org/10.5377/lrd.v45i1.19390

RESUMEN:

El apogeo de las redes sociales ha sido un factor preponderante para el establecimiento de un nuevo modelo de relaciones sociales, donde los seres humanos interactúan a través de espacios virtuales con características particulares de dinámica social, distintas a las formas convencionales de las últimas décadas. Por ello, estos espacios digitales forman parte central de las actividades del ser humano en el siglo XXI, dado que las personas han sido capaces de crear contenido de información, entretenimiento, realizar actos de comercio y mantener la comunicación con familiares y amistades.

No obstante, la irrupción del uso de sistemas de inteligencia artificial en las redes sociales para controlar o moderar el contenido en línea ha ocasionado que la restricción y censura al derecho a la libertad de expresión se erija como una nueva modalidad de vulneración a este derecho, particularmente cuando la supresión de contenidos se realiza por decisiones automatizadas por algoritmos; por ello, el objetivo del presente estudio se circunscribe a determinar si el uso de sistemas de inteligencia artificial para analizar publicaciones y decidir de forma automatizada su supresión o restricción, constituye un acto vulneratorio al derecho a la libertad de expresión.

PALABRAS CLAVE:

Derecho Constitucional, Derechos Fundamentales, Libertad de Expresión, Redes Sociales, Censura, Inteligencia Artificial, Decisiones Automatizadas.

Fecha de recepción: 31/08/24
Fecha de aprobación: 05/11/2024

1 Doctorando en derecho por la Universidad Nacional Autónoma de Honduras. Máster en derecho constitucional por la Universidad de Valencia, España. Abogado por la Universidad Nacional Autónoma de Honduras (UNAH). Asesor y Director en Juicio del Consultorio Jurídico Gratuito de la Facultad de Ciencias Jurídicas de la UNAH.
Correo electrónico: mauricio.cantor@unah.edu.hn.

2 Abogado in fieri por la Universidad Nacional Autónoma de Honduras. Miembro investigador del programa semillero de investigadores del Instituto de Investigación Jurídica de la UNAH. Correo Electrónico: gjgomezr@unah.hn

3 Abogada in fieri por la Universidad Nacional Autónoma de Honduras. Miembro investigador del programa semillero de investigadores del Instituto de Investigación Jurídica de la UNAH. Correo Electrónico: mnzuniga@unah.hn.

4 Abogado in fieri por la Universidad Nacional Autónoma de Honduras. Miembro investigador del programa semillero de investigadores del Instituto de Investigación Jurídica de la UNAH. Correo Electrónico: rssalinas@unah.hn.

5 Abogada in fieri por la Universidad Nacional Autónoma de Honduras. Miembro investigador del programa semillero de investigadores del Instituto de Investigación Jurídica de la UNAH. Correo Electrónico: lasanchezm@unah.hn.

FREEDOM OF EXPRESSION ON SOCIAL MEDIA: ANALYSIS OF AUTOMATED CENSORSHIP BY ARTIFICIAL INTELLIGENCE

Mauricio José C. Rosales[6]
Gerardo José Gómez Ramírez[7]
Marcia Nallely Zúniga Velásquez[8]
Richard Said Salinas Aguiriano[9]
Lorena Argentina Sánchez Maradiaga[10]
DOI: https://doi.org/10.5377/lrd.v45i1.19390

ABSTRACT:

The rise of social media has been a significant factor in establishing a new model of social relationships, where human beings interact through virtual spaces with unique dynamics that differ from conventional forms of the past decades. As a result, these digital spaces have become central to human activities in the 21st century, as people have been able to create content for information, entertainment, conduct commercial transactions, and maintain communication with family and friends.

However, the emergence of artificial intelligence systems in social media to control or moderate online content has led to a new form of violation of the right to freedom of expression, particularly when content suppression is carried out through algorithmic decisions. Therefore, the objective of this study is to determine whether the use of artificial intelligence systems to analyze posts and make automated decisions to suppress or restrict them constitutes a violation of the right to freedom of expression.

KEYWORDS:

Constitutional Law, Fundamental Rights, Freedom of Expression, Social Media, Censorship, Artificial Intelligence, Automated Decision.

Receipt date: 08/31/24
Approval date: 11/06/2024

6 PhD candidate in law at the National Autonomous University of Honduras. Master's degree in constitutional law from the University of Valencia, Spain. Lawyer from the National Autonomous University of Honduras (UNAH). Advisor and Director in Trial at the Free Legal Consultation Office of the Faculty of Legal Sciences at UNAH. Email: mauricio.cantor@unah.edu.hn.
7 Lawyer in progress at the National Autonomous University of Honduras. Research member of the research seed program at the Legal Research Institute of UNAH. Email: gjgomezr@unah.hn
8 Lawyer in progress at the National Autonomous University of Honduras. Research member of the research seedbed program at the Legal Research Institute of UNAH. Email: mnzuniga@unah.hn.
9 Lawyer in progress at the National Autonomous University of Honduras. Research member of the research seedbed program at the Legal Research Institute of UNAH. Email: rssalinas@unah.hn.
10 Lawyer in progress at the National Autonomous University of Honduras. Research member of the research seedbed program at the Legal Research Institute of UNAH. Email: lasanchezm@unah.hn.

I. INTRODUCCIÓN

El derecho a la libertad de expresión faculta al ser humano para manifestar sus ideas, opiniones, pensamientos, creencias e información sin temor a represalias, censura o restricciones ilegitimas por parte del Estado o de particulares. Este derecho posee una naturaleza jurídica bipartita: por un lado, en su dimensión individual, garantiza a cada persona la capacidad de comunicar libremente su punto de vista, información u opiniones; por otro, en su dimensión colectiva, protege el derecho de toda la sociedad a recibir, conocer y debatir tales puntos de vista, informaciones y opiniones sin interferencias que distorsionen u obstaculicen su acceso.

La evolución hacia la denominada web 2.0 o web social ha transformado la manera en que se ejerce este derecho. Las tecnologías de la información y la comunicación han permitido que los usuarios pasen de ser receptores pasivos de contenido a convertirse en participantes activos y creadores de información. Esto ha facilitado la interacción y colaboración entre usuarios, y ha dado lugar a oportunidades comerciales en plataformas interactivas de gran alcance, como Facebook, YouTube, Instagram, TikTok, LinkedIn y X (anteriormente Twitter). Estas plataformas han potenciado la libertad de expresión al permitir una mayor interactividad y visibilidad de las opiniones en un entorno globalizado.

En ese orden de ideas, este cambio de paradigma sociológico y tecnológico ha dado surgimiento a las denominadas redes sociales virtuales o en línea, las cuales se pueden definir de manera amplia como aquellos servicios de la sociedad de la información que ofrecen una plataforma de comunicación a través de internet y que permiten a los usuarios crear perfiles personales, interactuar y compartir contenido, como textos, fotografías, videos y enlaces, con otras personas dentro de la misma plataforma. Estas redes facilitan la creación de comunidades virtuales basadas en intereses comunes, relaciones personales o profesionales, permitiendo la conexión e interacción con otros usuarios por medio de mensajes privados, comentarios, publicaciones y reacciones.

Por otro lado, la irrupción de los sistemas de inteligencia artificial utilizados por las redes sociales, ha generado el surgimiento de nuevos desafíos y problemáticas para los derechos humanos, particularmente el derecho a la libertad de expresión, cuya problemática objeto del presente estudio se circunscribe a observar si el uso de sistemas de inteligencia artificial que analizan publicaciones para decidir unilateralmente suprimir o restringir el contenido de una publicación en una red social constituye una censura o vulneración al derecho a la libertad de expresión, sobre todo, cuando estas decisiones son de carácter automatizado.

Por ello, el objeto de estudio de la presente investigación se enfoca en el contenido esencial del derecho a la libertad de expresión en su dimensión individual; en consecuencia, el objetivo o propósito del estudio se circunscribe a determinar si el uso de sistemas de inteligencia artificial para analizar publicaciones y decidir de forma automatizada su supresión o restricción, constituye un acto vulneratorio al derecho a la libertad de expresión.

Lo anterior se justifica sobre la necesidad de analizar y abordar desde un enfoque jurídico doctrinario el impacto creciente que los sistemas de inteligencia artificial tienen sobre los derechos fundamentales en la era digital. La automatización en la toma de decisiones para suprimir o restringir publicaciones en plataformas de redes sociales

plantea serios desafíos éticos y jurídicos, ya que dichas decisiones pueden carecer del análisis contextual adecuado de las expresiones, sensibilidad y respeto a la pluralidad de voces y pensamientos. Además, el presente estudio adquiere relevancia al considerar que el derecho a la libertad de expresión no solo es un pilar democrático, sino que también es esencial para la construcción de una sociedad informada y plural. Por consiguiente, la posible vulneración de este derecho por sistemas automatizados subraya la necesidad de establecer límites y garantías claras que permitan un equilibrio entre la eficiencia en el uso de la tecnología y la protección de los derechos fundamentales.

II. METODOLOGÍA

Esta investigación es de tipo dogmática, basada en un modelo epistemológico de racionalismo jurídico y con un enfoque metodológico cualitativo, toda vez que el presente estudio descansa en la observación teórica del derecho a la libertad de expresión, su contenido esencial y alcances jurídicos, para así realizar una labor de análisis y raciocinio con el propósito de trasladar y acoplar las bases jurídicas del derecho a la libertad de expresión de la era analógica al ámbito tecnológico actual.

Todo lo anterior se fundamenta en los parámetros del método deductivo y procesos descriptivo, analítico y explicativo. Descriptivo, por cuanto se traza y detalla el contenido esencial del derecho a la libertad de expresión, su alcance y bienes jurídicos protegidos. Finalmente es analítico y explicativo, en virtud que se examinan estos componentes desarrollados del derecho a la libertad de expresión y se trasladan al ámbito digital para observar su impacto con los sistemas de inteligencia artificial utilizados por las redes sociales, procurando de esta manera, sustentar de forma clara y precisa, las razones que determinan la vulneración de los sistemas de inteligencia artificial en el derecho a la libertad de expresión.

III. DISCUSIÓN

1. El derecho a la libertad de expresión: Fundamentación y caracterización

La teoría del contenido esencial de los derechos fundamentales establece que cada derecho contiene un núcleo fijo, inmutable e irreductible que no puede ser desnaturalizado. Esto significa que cada derecho es constante y no se ve afectado o alterado por cambios en las circunstancias políticas, sociales, económicas o tecnológicas. Asimismo, dichas circunstancias no pueden implicar la modificación, eliminación o reducción de las facultades inherentes que cada derecho otorga, lo que garantiza la preservación de su significado y valor a lo largo del tiempo (López Sánchez, 2017).

Para definir el contenido esencial de un derecho fundamental, el Tribunal Constitucional Español considera que se deben seguir dos procedimientos metodológicos basados en el racionalismo jurídico. El primero, consiste en acudir a la naturaleza jurídica o modo de configurar dicho derecho, lo que requiere establecer una relación entre el lenguaje utilizado en las disposiciones normativas y el metalenguaje o convicciones generalmente admitidas entre los juristas, jueces y especialistas del derecho (Tribunal Constitucional Español, STC 11/1981, 1981). Esto implica que el intérprete constitucional deberá integrar en un proceso interpretativo el conjunto de términos, expresiones y estructuras lingüísticas del derecho en cuestión, y, al mismo tiempo, tomar en consideración el consenso jurídico en donde se describen, analizan, reflexionan y profundizan

jurídicamente las características básicas que definen inicialmente al derecho fundamental y, con ello, determinar su aplicación práctica.

Por otro lado, el segundo procedimiento busca identificar los intereses jurídicamente protegidos que constituyen el núcleo y centro de los derechos fundamentales. En este sentido, la esencia del contenido de un derecho se refiere a aquellos elementos que son absolutamente necesarios para que los intereses jurídicos que fundamentan el derecho sean protegidos de manera real, concreta y efectiva. Esto implica que el contenido esencial del derecho debe garantizar que estos intereses protegidos sean atendidos y resguardados de manera integral, asegurando así que el derecho conserve su plena efectividad y utilidad en la práctica jurídica (Tribunal Constitucional Español, STC 11/1981, 1981).

En ese orden de ideas, el derecho a la libertad de expresión, en general, otorga la facultad al ser humano de exteriorizar o difundir públicamente sus ideas, opiniones, pensamientos, creencias e información sin temor a represalias, censura o restricciones ilegitimas por parte del Estado o particulares (González Pérez, 2014). Por ello, se puede colegir que la naturaleza jurídica del contenido esencial de este derecho comprende una dimensión bipartida. Por un lado, este derecho abarca una esfera individual, en el sentido que otorga la facultad a toda persona para comunicar a otros el propio punto de vista y las informaciones u opiniones que se quieran y; por otro lado, la esfera colectiva, la cual comprende la facultad de todos los miembros en una sociedad a recibir y conocer tales puntos de vista, informaciones, opiniones, relatos y noticias, libremente y sin interferencias que las distorsionen u obstaculicen (Eduardo y Zelada, 2013).

Por tanto, dentro del derecho a la libertad de expresión convergen tres facultades de obrar esenciales, a saber: a) libertad de opinión, b) libertad de información y; c) libertad de prensa, las cuales constituyen la piedra angular para toda sociedad democrática y pluralista, en tanto y en cuanto permite el libre intercambio de ideas y opiniones, contribuyendo al debate público y al desarrollo de una sociedad informada y diversa (Eduardo y Zelada, 2013).

El derecho a la libertad de opinión comprende la protección del ámbito interno del individuo, es decir, el derecho a pensar y a formar juicios sin que nadie sea obligado a emitir su opinión, imponer una forma de pensamiento, o castigar a una persona por sus ideas; por otro lado, la libertad de información, refiere a la facultad de recibir y difundir información e ideas de toda índole, sin limitaciones por fronteras, a través de cualquier medio de comunicación, esto significa que todo ser humano puede recibir y difundir noticias y datos susceptibles de confirmación o, también, manifestar de forma escrita, oral o simbólica las creencias opiniones, proposiciones, peticiones, juicios valorativos, criticas o expresiones artísticas, indistintamente del medio a través del cual se exteriorizan. Finalmente, la libertad de prensa consiste en la protección especializada que tienen los medios de comunicación para no recibir restricciones o censuras indebidas (Eduardo y Zelada, 2013).

Atendido lo anterior, tradicionalmente los derechos fundamentales han sido entendidos como normas que regulan las relaciones entre el Estado y los individuos, imponiendo límites exclusivamente al poder estatal para proteger la libertad y la dignidad de las personas. Sin embargo, a partir de la doctrina alemana de la drittwirkung, se transformó el paradigma de

las relaciones jurídicas derivadas de derechos fundamentales, en tanto y en cuanto las prerrogativas esenciales, mínimas e inherentes al ser humano también son oponibles frente a los actos que ejecuten particulares, en virtud que los derechos fundamentales irradian en el resto del ordenamiento jurídico; por consiguiente, las obligaciones generales de respeto y garantía que inicialmente solamente se le exigía al Estado, también deben ser cumplidas por los particulares en las relaciones interindividuales (Mora Sifuentes, 2017).

Por tanto, el derecho a la libertad de expresión en su dimensión individual se vulnerará cuando autoridades estatales o particulares, por acción u omisión, generen actos constitutivos de censura previa, directa o indirecta. Así pues, la censura previa se refiere a la prohibición o revisión de contenidos antes de que se publiquen o difundan, esto significa que las autoridades o los particulares revisan la información antes de su divulgación y pueden impedir que se publique si consideran que atentan contra ciertos intereses; por otro lado, la censura directa se refiere a aquellos actos tendientes a eliminar o suprimir explícitamente contenidos considerados inapropiados, peligrosos o contrarios a ciertos intereses, los cuales se pueden manifestar a través de emisión de legislaciones restrictivas, bloqueos de contenido, represión de medios de comunicación y; la censura indirecta, consiste en el establecimiento de un ambiente en el que es difícil o arriesgado para los medios de comunicación o individuos expresar libremente sus ideas (Eduardo & Zelada, 2013).

1.1. Límites al derecho a la libertad de expresión.

La Corte Interamericana de Derechos Humanos ha establecido que el derecho a la libertad de expresión no absoluto, por lo que puede estar sujeto a ciertas restricciones que, para ser legítimas, deben cumplir los siguientes requisitos: a) **idoneidad,** que implica que la medida restrictiva debe ser capaz de contribuir al logro de la finalidad establecida; b) **legalidad,** que exige que toda restricción esté claramente definida en una ley previa, comprensible y accesible para los ciudadanos; c) **necesidad,** que implica que la medida adoptada sea indispensable y que no exista una alternativa menos restrictiva; d) **proporcionalidad,** que establece que la restricción debe mantener una relación razonable entre los beneficios que genera y los derechos que limita; y e) **finalidad legítima,** que significa que las restricciones deben perseguir un objetivo compatible con la protección de los derechos humanos (Corte IDH. Caso Jenkins Vs. Argentina, 2019).

En el ámbito digital y, particularmente en internet, se ha considerado que los estándares generales para restringir derechos deben ser evaluados con una perspectiva de sistémica digital. Esta refiere a la comprensión del espacio digital como un ecosistema complejo, donde interactúan factores técnicos, sociales y normativos, que no puede abordarse con los mismos enfoques tradicionales, formular enfoques alternativos y específicos para la imposición de restricciones a la libertad de expresión en internet, que se adapten a sus características singulares, y que a la vez reconozcan que no deben establecerse restricciones especiales a contenido de los materiales que se difunden a través de internet (Botero, 2012).

Sin embargo, en casos excepcionales, cuando se está frente a contenidos abiertamente ilícitos o a discursos no resguardados por el derecho a la libertad de expresión, como, por ejemplo, la propaganda de guerra y la apología del odio que constituya incitación a la violencia, la incitación

directa y pública al genocidio, y la pornografía infantil, resulta admisible la adopción de medidas obligatorias de bloqueo y filtrado de contenidos específicos. En estos casos, la medida debe someterse a un estricto juicio de proporcionalidad y estar cuidadosamente diseñada y claramente limitada de forma tal que no alcance a discursos legítimos que merecen protección (Botero Marino, 2013)

Por tanto, en el ámbito digital, el derecho a la libertad de expresión encontrará su límite en el derecho de los demás, por lo que las medidas de restricción deben contar con salvaguardas que eviten el abuso, como la transparencia respecto de los contenidos cuya remoción haya sido ordenada y que los sistemas de filtrado de contenidos impuestos por proveedores de servicios comerciales que no sean controlados por el usuario final constituyen una forma de censura previa y no representan una restricción justificada a la libertad de expresión, por lo que debe exigirse que los productos destinados a facilitar el filtrado por los usuarios finales estén acompañados por información clara dirigida a dichos usuarios acerca del modo en que funcionan y las posibles desventajas si el filtrado resulta excesivo.

2. La relación sinalagmática entre el derecho a la libertad de expresión y las redes sociales

La noción de redes sociales en la actualidad es un término utilizado popularmente para referirse en lato sensu a todas aquellas plataformas digitales en línea en donde los usuarios crean, comparten y difunden contenido; no obstante, tal concepción no se limita exclusivamente al ámbito digital, en virtud la conceptualización de red social surge en la primera mitad del siglo XX a través de las aportaciones del psicólogo y filósofo alemán Kurt Lewin, quien sostenía que la percepción, la conducta de los individuos y la estructura misma del grupo en el que se encuentran inmersos, están inscritos a un espacio social formado por el grupo y el entorno que lo rodea, constituyendo de esa manera un campo de relaciones (Avila Toscano & Maradiaga Orozco, 2012).

Por ello, una red social puede ser entendida como la vinculación de un conjunto de actores por medio de relaciones sociales definidas; en consecuencia, la idea de red aplicada al contexto de las relaciones entre sujetos implica, per se, la existencia de una estructura social conformada por individuos, desde la cual de manera directa o indirecta se encuentran unidos gracias al ejercicio de compartir diversas interacciones surgidas espontánea o intencionalmente, las que además están medidas por un patrón social que determina la forma como se intercambian recursos.

Con la evolución a la denominada web 2.0 o web social, las tecnologías de la información y comunicación han permitido que los usuarios interactúen y colaboren entre sí, pasando de ser sujetos pasivos, en donde simplemente recibían la información en la red o la publicaban sin ninguna posibilidad de interacción, a convertirse realmente en sujetos activos, creadores de contenido y con ello teniendo la posibilidad de entablar relaciones sociales, crear contenido y ejecutar negocios comerciales orientado hacia grandes volúmenes de usuarios en la red, lo cual se debe, principalmente, al surgimiento de aplicaciones o plataformas que usan una interface con altos niveles de interactividad, tales como: Facebook, Youtube, Instagram, Tik Tok, LinkedIn y Twitter (ahora X).

En ese orden de ideas, este cambio de paradigma sociológico y tecnológico ha dado surgimiento a las denominadas redes sociales

virtuales o en línea, las cuales se pueden definir de manera amplia como aquellos servicios de la sociedad de la información que ofrecen a los usuarios una plataforma de comunicación a través de internet para que estos generen un perfil con sus datos personales, facilitando la creación de redes en base a criterios comunes y permitiendo la conexión e interacción con otros usuarios (Ortiz López, 2013). De manera tal que, esta interacción estará marcada por algunos aspectos particulares como el anonimato total o parcial de los usuarios, el contacto sincrónico o asincrónico y la seguridad o inseguridad que dan las relaciones que se suscitan por esta vía.

Por ello, las características que rigen las modalidades del derecho a la libertad de expresión en las redes sociales se caracterizan por las manifestaciones del lenguaje hablado, escrito y el soporte visual, independientemente del medio utilizado, ya que en cada publicación, comentario, imagen o video, existen rasgos característicos de expresión oral, en virtud que se presentan las marcas propias de la oralidad, tales como la expresión a través del uso de palabras en mayúsculas, interjecciones, repetición de signos de puntuación, así como también la presencia de los elementos que hacen referencia al lenguaje corporal en forma de emoticonos, los cuales suplen las expresiones faciales y actitudinales del hablante en una conversación (Ortiz López, 2013).

Por otro lado, en relación con la característica de la comunicación escrita, puede inferirse que estas plataformas brindan la posibilidad al usuario de convertirse en autor a través de las publicaciones colgadas en los perfiles personales, comentarios o los denominados tweets. A diferencia de los textos tradicionales, la escritura producida en el ámbito virtual es mucho más mudable, dinámica y la información es mucho más volátil, ya que implican la interacción constante de varios participantes que añaden, borran o modifican su contenido.

Asimismo, las características de expresividad en las redes sociales adquiere una serie de elementos propios solamente del lenguaje de las redes sociales virtuales, tal sería el caso del soporte visual, el cual se manifiesta a través de fotografías, vídeos, ficheros de audio, etc., los efectos gráficos, como los emoticonos, emojis, imágenes en formato gif (configuración de imagen en movimiento) y los memes, popularmente asociados al humor de imagen y texto y; finalmente, la presencia de hipervínculos, manifestado por medio de un link o hashtag que realizan la conexión con la información de otra página web y que ofrece la posibilidad de la intertextualidad (Ortiz López, 2013).

Por tanto, se puede colegir que las redes sociales solo son otro medio de comunicación aceptado por la sociedad en general, por lo que las prerrogativas y contenido esencial del derecho a la libertad de expresión será aplicable también al ámbito de las redes sociales, por lo que las obligaciones de respeto y garantía que corresponden a este derecho, deberán también ser cumplidas y salvaguardadas por las empresas propietarias de redes sociales y; en consecuencia, abstenerse a cualquier conducta, por acción u omisión, que genere censura previa, directa o indirecta.

3. La inteligencia artificial: funcionalidad en las redes sociales e impactos en el derecho a la libertad de expresión

La inteligencia artificial puede ser definida como los sistemas de software o hardware diseñados por humanos que, en función de un objetivo complejo, actúan en la dimensión física o digital percibiendo su entorno a través de la obtención,

interpretación, razonamiento y procesamiento de datos con el propósito de tomar las mejores decisiones para lograr el objetivo propuesto. En el ámbito de las redes sociales, la inteligencia artificial ha jugado un rol sustancial para mejorar la experiencia de los usuarios en línea. En ese sentido, ha sido sujeta a una pluralidad de funciones entre las cuales se encuentran el análisis de emociones, generación de contenido, recomendaciones personalizadas, detección de contenido inapropiado, chatbots, reconocimiento y creación de imágenes y publicidad dirigida (Cabrera, Ketty, León-Alberca y Arpi, 2024).

Por ello, uno de los principales sistemas de inteligencia artificial que funcionan dentro de las redes sociales son aquellas destinadas a la detección de contenido catalogado como inapropiado. Este tipo de inteligencia artificial, también denominada de moderación de contenido, son sistemas que clasifican el contenido generado por el usuario con base en la coincidencia (matching) o en la predicción (classification), tomando una decisión que tiene como resultado la eliminación, bloqueo o suspensión de la cuenta, entre otros (Gorwa, Binns y Katzenbach, 2020).

Las IA de moderación de contenido pueden clasificarse según su funcionamiento y sus consecuencias. Según su funcionamiento, la IA de moderación de contenido suele ser de matching o de classification. Las IA de matching implican el hashing, es decir, el proceso de transformar un ejemplo conocido de una pieza de contenido en una cadena de datos destinada a reconocer contenido idéntico. Sin embargo, una de las problemáticas principales que enfrenta el hashing se circunscriba a que, si el contenido analizado tiene una ligera perturbación o diferencia con el ejemplo de la base de datos, puede pasar desapercibido por el sistema de IA. Por esta razón, se han creado sistemas de hashing que buscan no identificar coincidencias exactas entre el ejemplo de la base de datos y el contenido analizado, sino similitudes (Gorwa, Binns y Katzenbach, 2020).

Por otro lado, las IA de classification son distintas a las mencionadas previamente. En lugar de necesitar una pieza de contenido de la base de datos para identificar contenido exacto o similar, las IA de classificación analizan contenido que no corresponde con una versión previa en la base de datos, es decir, detectan patrones en los datos recibidos. En la actualidad, esta es la IA de moderación de contenido más utilizada en las redes sociales (Gorwa, Binns y Katzenbach, 2020).

Atendido lo anterior, se puede colegir que la IA ha adquirido diversas funcionalidades en el contexto de las redes sociales, siendo una de la más importantes el rol de moderadora de contenido. Sin embargo, la moderación del contenido a cargo de la IA, si bien puede significar reducción de costos, mayor cobertura de contenido, eficacia y rapidez, también puede tener consecuencias negativas en el ejercicio y goce del derecho a la libertad de expresión, como falsos positivos y falsos negativos, los sesgos y discriminación algorítmica, el procesamiento a gran escala de datos personales y perfilamiento, la censura previa, la supervisión inadecuada, la falta del debido proceso y la rendición de cuentas y transparencia (Llansó, Hoboken, Leersen y Harambam, 2020).

Por esta razón, el Derecho Internacional de los Derechos Humanos ha redoblado esfuerzos para crear estándares a fin de regular el uso de sistemas de inteligencia artificial y el impacto que puedan tener en el derecho a la libertad de expresión. Según las Directrices Éticas para una IA Fiable, la IA debe someterse a los siguientes estándares: i) acción y supervisión humana, ii) solidez técnica y seguridad, iii) gestión de la privacidad y los datos, iv) transparencia, v) no discriminación

y; vi) rendición de cuentas (Grupo independiente de expertos de alto nivel sobre inteligencia artificial, 2019).

Bajo esa premisa, la acción y supervisión humana implica realizar evaluaciones de impacto sobre los derechos humanos antes de su implementación e incluir una evaluación de las posibilidades de reducir esos riesgos o de justificarlos como necesarios en una sociedad democrática, así como garantizar el derecho a no ser objeto de una decisión basada exclusivamente en procesos automatizados cuando tal decisión produzca efectos jurídicos o les afecte de forma similar. Asimismo, implica la supervisión de la IA a través de la participación y el control humano, garantizando, a su vez, la supervisión por parte de funcionarios públicos.

La solidez técnica y seguridad asegura la protección de la IA frente a sus vulnerabilidades, evitando que agentes malintencionados perturben o le hagan daño al sistema. De igual manera, garantiza que la IA funcione adecuadamente en distintos contextos y que se comporte de la misma manera en las mismas condiciones. Por otro lado, la gestión de la privacidad y datos hace referencia a que se debe garantizar el derecho a la privacidad de los titulares de datos personales gestionados por la IA.

La transparencia supone la trazabilidad del sistema, es decir, los motivos o razones por las cuales llega a una decisión determinada, y también la explicabilidad del sistema, esto es, que las decisiones de la IA sean comprensibles para el humano y que se tenga la posibilidad de rastrearlas. Por su parte, el cumplimiento de la no discriminación comprende la supervisión de las decisiones de la IA para identificar sesgos o prejuicios y asegurar que personas de diversos contextos participen en el desarrollo e implementación de esta. Por último, la auditabilidad conlleva la capacidad para evaluar algoritmos y la garantía de mecanismos para obtener una reparación en caso de efectos adversos.

En el caso de la acción y supervisión humana, es menester hacer énfasis en las decisiones automatizadas. Las decisiones automatizadas son aquellas realizadas por un sistema de IA sin necesidad de intervención humana. En los últimos años, este tipo de decisiones han sido cuestionadas por el potencial impacto que tienen sobre los derechos humanos y han despertado el debate sobre el derecho a no ser objeto de decisiones automatizadas, también llamado como el "derecho al control humano". Este derecho implica la capacidad de que intervengan seres humanos durante el ciclo de diseño del sistema de inteligencia artificial y en el monitoreo de su funcionamiento, con el fin de evitar un impacto negativo en los derechos humanos, esto significa que esta prerrogativa se constituye como un corolario que refuerza las manifestaciones de todos los derechos fundamentales en el plano digital, en particular el derecho a la libertad de expresión (Eguíluz, 2020).

En ese orden de ideas, para identificar la existencia de una decisión automatizada, deben concurrir los siguientes requisitos, a saber: i) una decisión ii) basada únicamente en procesamiento automatizado que incluya la elaboración de perfiles, iii) y que produzca efectos jurídicos o consecuencias similares. El primer requisito hace referencia a que un comportamiento particular ha sido tomado respecto de una persona y tiene un efecto vinculante. El segundo requisito implica que la decisión final, si bien le corresponde a una persona humana, carece de una contribución sustancial de su parte previo a la formalización. En este mismo requisito, se impone como condición la existencia de tratamiento automatizado que incluya la elaboración de perfiles, lo cual se

verifica cuando el tratamiento automatizado persigue evaluar determinados aspectos personales. Por último, el tercer requisito establece que la decisión debe determinar derechos o deberes de la persona o impactar su bienestar (Mendoza y Bygrave, 2017).

Frente a la presencia de una decisión en los términos previamente descritos, el derecho al control humano se manifiesta a través de los derechos a la explicación y a la impugnación de la decisión automatizada (Malgieri, 2019). El derecho a la explicación tiene una naturaleza bifronte, es decir, puede ser ex ante facto y ex post facto. El derecho de explicación ex ante tiene lugar antes de la decisión automatizada y hace referencia a informar acerca de la lógica involucrada, consecuencias previstas y funcionalidad en general del sistema de decisiones automatizado. El derecho de explicación ex post tiene lugar después de la decisión automatizada y versa sobre las razones y circunstancias individuales que conllevaron a un sistema de decisiones automatizado a tomar determinada decisión (Wachter, Mittelstadt y Floridi, 2017). Por último, el derecho a la impugnación implica un procedimiento contencioso que pretende objetar la decisión automatizada (Brkan, 2018, p. 13).

El derecho a la explicación y el derecho a la impugnación guardan una estrecha relación. Por un lado, la impugnación de una decisión automatizada se vuelve inefectiva sin saber la funcionalidad general de la IA y las razones y circunstancias individuales que la llevaron a tomar tal decisión, dado que son elementos necesarios para formular alegatos y garantizar la defensa de los intereses del titular de datos. Por otro lado, el derecho de explicación carece de un potencial efecto vinculante y reparador de derechos sin la presencia de un recurso destinado a anular la decisión automatizada y obtener adecuada intervención humana en pro de la defensa de los intereses del titular de los datos. En consecuencia, ambos derechos son imprescindibles para la tutela de derechos humanos, especialmente de la libertad de expresión, frente a las decisiones automatizadas.

3.1. Las decisiones automatizadas de los sistemas de inteligencia artificial y sus impactos en el derecho a la libertad de expresión.

Las decisiones de la inteligencia artificial tienen un impacto en la libertad de expresión a través de la censura algorítmica, ya sea de forma previa, directa o indirecta. La censura algorítmica previa, entendida como la revisión y posible prohibición de un contenido antes de que se haga público, es una práctica prohibida absolutamente en el marco del artículo 13 de la Convención Americana sobre Derechos Humanos, admitiendo como excepción los espectáculos públicos con el fin de regular el acceso de la infancia y la adolescencia a ellos, así como los contenidos sensibles o lesivos a la dignidad humana (Eduardo & Zelada, 2013). Por lo tanto, las decisiones automatizadas que censuran previamente el contenido antes que se publique constituyen, por sí mismas, una vulneración al derecho a la libertad de expresión.

La censura directa e indirecta de la libertad de expresión suelen manifestarse a través de la eliminación de contenido, aplicación de faltas, restricción de cuentas, inhabilitación de cuentas, eliminación de páginas y grupos, reducción de la distribución del contenido problemático y proporcionar contexto sobre contenido delicado o engañoso. Este tipo de impedimentos que controlan el flujo de contenido después de realizada su publicación implica una restricción al derecho a la libertad de expresión, sin embargo, la justificación sobre si la restricción es legítima o no cons-

tituye el punto de inflexión para determinar si hay una vulneración al derecho o no (Botero Marino, 2013). En ese sentido, habrá una vulneración al derecho a la libertad de expresión cuando a la persona objeto de la decisión automatizada no se le garantice el derecho al control humano, manifestado a través de la explicación de la decisión y su impugnación.

En síntesis, la censura algorítmica previa, por sí misma, vulnera el derecho a la libertad de expresión. En cambio, la censura directa e indirecta no pueden por sí solos vulnerar este derecho, sino que debe haber una falta de explicación de la funcionalidad del algoritmo y de las razones que motivaron su decisión, así como una falta de acceso a un medio de impugnación de la decisión automatizada.

3.2. Funcionalidad de los sistemas de inteligencia artificial en las redes sociales

Las plataformas digitales, que facilitan el acceso a información y el debate en línea, han dado origen a un internet abierto. Pese a todo, con el avance tecnológico y el aumento de fenómenos como la difusión de desinformación, los discursos de odio y la vigilancia digital, se ha vuelto más complejo aplicar las arcaicas herramientas y estándares para proteger la libertad de expresión (Lanza y Jackson, 2021). Es en este contexto, que surge la necesidad del uso de la IA como herramienta idónea por su capacidad de procesar masivas cantidades de datos e identificar patrones y temas. En el marco de las redes sociales, la IA mejora la experiencia de los usuarios facilitando la moderación de contenidos, con el fin de mitigar posibles discursos de odio, discriminatorios, terroristas, etc. (Calva-Cabrera, León-Alberca y Arpi-Fernández, 2024).

Por ello, la utilización de sistemas de inteligencia artificial para revisar y moderar el contenido en las redes sociales ha dado el surgimiento a nuevas manifestaciones de vulneración al derecho a la libertad de expresión; en consecuencia, resulta oportuno observar las políticas y formas de revisión y monitoreo de publicaciones para observar las publicaciones que se realizan en Facebook y X (Antes Twitter), constituye un sistema de decisión automatizada que pueda vulnerar el derecho a la libertad de expresión.

a. X (Twitter).

La red social X destaca que son de interés público y, por tanto, menos propensos a ser eliminados aquellos tuits que emiten funcionarios gubernamentales pues es importante conocer qué hacen con el fin de debatir sus acciones u omisiones. Así, se da prevalencia a la difusión de contenido de interés público basándose en los siguientes tres criterios que conforman una excepción a la remoción directa de contenido: i) El tuit incumple alguna o varias reglas de X, ii) El autor tiene una cuenta de alto perfil y iii) La cuenta representa a un integrante actual o potencial de Gobierno o Poder legislativo local, nacional o supranacional o un candidato presidencial.

En el caso anterior, X agrega un aviso para contextualizar la violación y limitar las interacciones con el tuit, como los "Me gusta" y "Retweets". Esto busca restringir su alcance, pero permite al público verlo y debatir sobre el tema. En estos casos, el equipo de "Trust & Safety" de X realiza un segundo análisis para decidir si conservar o no la visibilidad del tuit, basándose en criterios de interés público. Las recomendaciones iniciales del equipo son revisadas por un grupo interno de expertos antes de que los líderes de "Trust & Safety" tomen la decisión final sobre si aplicar el aviso o eliminar el tuit (Twitter, 2024).

Por otro lado, con relación al contenido modificado o falsificado, X no especifica cómo concluye que podría causar confusión, engaño o, si intervienen profesionales en este análisis, como ocurre con los contenidos de interés público. De esa manera, la falta de precisión podría dilucidar que X emplearía IA a los fines de revisar: i) incluir elementos multimedia que hayan sido considerablemente alterados, manipulados o falsificados con la intención deliberada de engañar; o ii) incluir elementos multimedia que se comparten de manera engañosa o con un contexto falso e; iii) incluir elementos multimedia que puedan provocar confusión generalizada con respecto a asuntos públicos, afectar la seguridad pública o provocar daños graves (Twitter, 2024).

En síntesis, se puede colegir que las medidas automatizadas que Twitter adopta ante un contenido que la misma plataforma califica como falso o alterado, varían dependiendo de la gravedad en el impacto de publicación, obteniéndose una penalización que va desde la disminución de la amplificación e interacción con la publicación hasta la eliminación del post. En el primer caso, se etiqueta la publicación como engañosa o fuera de contexto y los usuarios obtienen un mensaje previo a leer o compartir la publicación. No obstante, la información puede ser contrastada mediante las notas que la misma comunidad de Twitter añade sobre el contexto de la información publicada.

b. Facebook

Facebook informa en su plataforma que su estrategia para detener la información errónea consiste en tres acciones puntuales: i) Exigir el estricto cumplimiento de las políticas, ii) aplicar el aprendizaje automático a fin de asistir al equipo de respuesta para que detecten información errónea y apliquen las políticas contra las cuentas no auténticas y; iii) actualizar la detección de cuentas falsas en Facebook, lo que obstaculiza el envío masivo de spam. Además, se automatiza la clasificación de noticias falsas (Meta, 2024).

Para la detección de información errónea, Facebook cuenta también con verificadores de datos independientes, certificados por la Red Internacional de Verificación de Datos. Cuando estos verificadores califican un contenido como falso, Facebook lo clasifica más bajo en el News Feed, reduciendo su visibilidad en un 80%. Este cambio ha llevado a Facebook a pasar de depender de las denuncias de los usuarios a una moderación proactiva y automatizada de contenido (Larrondo & Grandi, 2021).

En ese orden de ideas, gran parte de las estrategias de revisión de contenido están automatizadas por algoritmos de IA. En respuesta a este problema, nace un mecanismo de revisión independiente y descentralizado denominado Facebook Oversight Board (FOB), el cual opera modelando las políticas y estándares de la libertad de expresión en línea. Así, Facebook, Instagram o cualquiera de sus usuarios pueden remitir un caso de revisión al FOB por dos causales: i) por reclamos por la restitución de publicaciones que hubieran sido removidas por la plataforma y, ii) por reclamos para dar de baja contenido que los usuarios consideren que debería ser eliminado. Para recurrir al FOB, los usuarios primero deberán agotar los mecanismos de revisión internos de Facebook (Lanza & Jackson, 2021, p. 10).

Por tanto, se puede colegir que los sistemas de IA están presentes tanto en la distribución de contenido en las redes sociales digitales, como en las restricciones a las publicaciones y comentarios que se realizan en dichas plataformas. En su mayoría, las interacciones de los usuarios es-

tán limitadas por la automatización en la moderación del contenido que autoriza a la red social aplicar diversos mecanismos de censura previa, lo que lleva a restringir la libertad de expresión si no se contextualiza debidamente la expresión y si no existe control humano que garantice medios para impugnar una decisión automatizada.

IV. CONCLUSIONES

La libertad de expresión, en tanto defiende el derecho a buscar, comunicar libremente ideas y de recibirlas, ha encontrado en las redes sociales un nuevo desafío que afrontar. En primer término, porque el actor encargado de la regulación del flujo de ideas no es el Estado, sino particulares, lo que aviva el debate sobre el efecto horizontal de los derechos humanos y; en segundo término, porque las tecnologías involucradas en la aplicación de mecanismos de censura no son convencionales; son herramientas capaces de procesar vastas cantidades de datos en un tiempo récord y que prescinden de la intervención humana.

No obstante, el derecho a la libertad de expresión no es un derecho estático, sino dinámico, pues en la medida que la tecnología avanza y el mundo se transforma, también se transforma el contenido esencial de este derecho. En ese sentido, la libertad de expresión no ha quedado limitada a la realidad análoga, sino que se ha extendido a la realidad digital de la cual las redes sociales participan y se han convertido en espacios por excelencia para el debate público. Por tanto, la libertad de expresión veda, en el ámbito de las redes sociales, la censura previa y todo tipo de impedimento u obstáculo a la búsqueda, difusión y recepción de expresiones, ideas u opiniones por parte de los usuarios de dichas plataformas.

En esta tesitura, la libertad de expresión se ve vulnerada con las decisiones automatizadas a través de la censura algorítmica previa y el resto de los impedimentos y obstáculos a la libertad de expresión practicados por las redes sociales. La censura algorítmica previa constituye, por sí misma, una vulneración al derecho a la libertad de expresión, en tanto existe dentro del Derecho Internacional de los Derechos Humanos una prohibición absoluta de esta práctica que debe ser rechazada en el derecho interno de los Estados.

Por otro lado, la presencia de obstáculos al derecho a la libertad de expresión restringe o limitan el contenido de este derecho, pero ello no es causa suficiente para establecer la vulneración del mismo, toda vez que es menester verificar si tal restricción fue justificada de conformidad con los estándares del test de restricción de derechos o porque los discursos o expresiones en disputa están fuera de la órbita de protección del derecho. En consecuencia, la restricción al derecho a la libertad de expresión será injustificada y, consecuentemente, comportará una violación al derecho, cuando no se garantice el control humano sobre las decisiones automatizadas, siendo este el remedio jurídico idóneo para tal fin.

Por esta razón, se vuelve indispensable la creación de estándares y regulaciones para la inteligencia artificial utilizada en el contexto de las redes sociales, a fin de garantizar el goce y ejercicio del derecho a la libertad de expresión. Sin limitaciones legales al uso de la inteligencia artificial, la censura, falta de transparencia y proliferación del sesgo por automatización hacen imposible el goce y ejercicio de este derecho.

V. BIBLIOGRAFÍA.

Álvarez, D. (2004). *Libertad de expresión en internet y el control de contenidos ilícitos y nocivos.* Centro de estudios en derecho informático de la Facultad de Derecho de la Universidad de Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/107436/alvarez_d.pdf?sequence=3&isAllowed=y

Avila Toscano, J. H., & Maradiaga Orozco, C. (2012). Redes Sociales: Un ejercicio caracterológico. En J. H. Avila Toscano, Redes sociales y análisis de redes. *Aplicaicones en el contexto comunitario y virtual* (págs. 14-47). Colombia: Azul y violeta editores.

Botero Marino, C. (2013). *Informe de la relatoría especial para la libertad de expresión.* Washington D.C.: Comisión Interamericana de Derechos Humanos.

Botero, C. (2012). *Informe de la relatoría especial para la libertad de expresión.* Washington D.C.: Comisión Interamericana de Derechos Humanos.

Calva-Cabrera, K. D., León-Alberca, T., & Arpi-Fernández, C. G. (2024). Inteligencia Artificial en las redes sociales digitales. Espejo de Monografías de Comunicación Social, (23), 15-35.

Corte IDH. Caso Jenkins Vs. Argentina, Serie C No. 397 (Corte Interamericana de Derechos Humanos 26 de Noviembre de 2019).

Eduardo, B., & Zelada, C. J. (2013). Artículo 13. Libertad de pensamiento y de expresión. En C. Steiner, & P. Uribe, *Convención Americana sobre Derechos Humanos Comentarios* (págs. 320-342). Bogotá: Konrad Adenauer.

Eguíluz, J. (2020). *Desafíos y Retos que Plantean las Decisiones Automatizadas y los Perfilados para los Derechos Fundamentales.* Revista Estudios de Deusto, vol. 68/2, 325-367.

Meta, Inc. 2024. Condiciones del servicio. En: Facebook [en línea]. Disponible en: https://www.facebook.com/legal/terms [consulta: 17 de junio 2024].

González Pérez, L. R. (2014). Libertad de expresión. En E. F. Mac-Gregor, F. Martínez Ramírez, & G. Figueroa Mejía, *Diccionario de derecho procesal constitucional y convencional* (págs. 885-887). Ciudad de México: Universidad Nacional Autónoma de México. Instituto de Investigaciones Jurídicas. Poder Judicial de la Federación, Consejo de la Judicatura Federal.

Gorwa, R; Binns, R; Katzenbach, C. (2020). *Algorithmic content moderation: technical and political challenges in the automation of platform governance.* Big Data and Society, 1-15.

Grupo independiente de expertos de alto nivel sobre inteligencia artificial. (2019). *Directrices éticas para una IA fiable.* Bruselas: Comisión Europea.

Lanza, E. & Jackson, M. (2021). *Moderación De Contenidos Y Mecanismos De Autorregulación: El Oversight Board De Facebook Y Sus Implicancias Para América Latina.* Diálogo Interamericano.

Llansó, E; Hoboken, J; Leersen, P; Harambam, J. (2020). *Artificial Intelligence, Content Moderation and Freedom of Expression.* Transatlantic Working Group. Recuperado de: https://www.semanticscholar.org/paper/Artificial-Intelligence%2C-Content-Moderation%2C-and-of-Llanso/632f632a79994254506a2d017fd5ca804b6067cc

Malgieri, G. (2019). *Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations.* Elsevier Ltd. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0267364918303753?ref=pdf_download&fr=RR-2&rr=8bb61af26d123ba7

Mendoza, I y Bygrave, L. (2017). The right not to be subject to automated decisions based on profiling. Springer International Publishing. *EU Internet Law, 77-98*.

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B. (2020). *AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy for the AI landscape.* Publications Office of the European Union

Wachter, S; Mittelstadt, B y Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law,* 76-99.

López Sánchez, R. (2017). Indeterminación y contenido esencial de los derechos humanos en la constitución mexicana. C*uestiones Constitucionales,* 229-263.

Mora Sifuentes, F. M. (2017). La influencia de los derechos fundamentales en el ordenamiento: su dimensión objetiva. *Boletín mexicano de derecho comparado,* 1215-1258.

Ortiz López, P. (2013). Redes sociales: funcionamiento y tratamiento de información personal. En A. Rallo Lombarte, & R. Martínez Martínez, Derecho y redes sociales (págs. 117-144). Madrid: Civitas Thomson Reuters.

STC 11/1981 (Tribunal Constitucional Español 8 de Abril de 1981).

Twitter, inc. 2024. Twitter términos de servicio. En: *twitter* [en línea]. Disponible en: https://twitter.com/es/tos [consulta: 15 de junio 2024]