Ivs Fvgit, 29, 2026, pp. 241-268
ISSN: 1132-8975

LA TRANSPARENCIA ALGORÍTMICA EN LA MODERACIÓN DE CONTENIDOS EN REDES SOCIALES

María Sáez de Propios
Universidad Católica de Ávila
maria.saezdepropios@ucavila.es

RESUMEN:

La moderación de contenidos necesita la implementación de sistemas de transparencia. Este texto analiza los desafíos de la moderación de contenidos en redes sociales y el papel de la Inteligencia artificial (IA) en este proceso. Las empresas deben mejorar la transparencia de sus algoritmos para asegurar una mayor neutralidad, y proporcionar información clara sobre su funcionamiento y personalización de contenidos. Aunque la IA se usa en la autorregulación, la falta de intervención humana plantea preocupaciones sobre censura y limitación de la libertad de expresión. Este trabajo evalúa el equilibrio entre seguridad y libertad de expresión, así como la transparencia y responsabilidad algorítmica de plataformas como Facebook, Twitter (ahora X) e Instagram. Mediante una investigación cualitativa, se han identificado problemas y fallos en la moderación, por lo que se evidencia la importancia de una implementación adecuada de la IA y el aprendizaje automático.

PALABRAS CLAVE:

Transparencia, autorregulación, moderación, algoritmos, inteligencia artificial.

ALGORITHMIC TRANSPARENCY IN SOCIAL MEDIA CONTENT MODERATION

ABSTRACT:

Content moderation requires the implementation of transparency systems. This text analyses the challenges of content moderation in social networks and the role of Artificial Intelligence (AI) in this process. Companies need to improve the transparency of their algorithms to ensure greater neutrality, and provide clear information about how they work and how they personalise content. Although AI

is used in self-regulation, the lack of human intervention raises concerns about censorship and limitation of freedom of expression. It assesses the balance between safety and freedom of expression, as well as the transparency and algorithmic accountability of platforms such as Facebook, Twitter (now X) and Instagram. Through qualitative research, problems and failures in moderation have been identified, highlighting the importance of proper implementation of AI and machine learning.

KEYWORDS:
Transparency, self-regulation, moderation, algorithms, artificial intelligence.

1. INTRODUCCIÓN

El uso de la inteligencia artificial (IA) en la moderación de contenidos en las redes sociales surge como una solución prometedora ante los desafíos crecientes en este ámbito. La enorme cantidad de datos generados, la persistencia de las infracciones y la necesidad de reducir la intervención humana en la toma de decisiones han impulsado el desarrollo de herramientas que las plataformas de redes sociales han implementado para automatizar la moderación de su contenido. La magnitud del contenido que día a día se genera en las plataformas como Facebook, Instagram o X (antiguo Twitter) ha provocado la necesidad de crear estos sistemas que puedan dar respuesta a una moderación de contenidos prácticamente en tiempo real. Por tanto, la IA en la moderación de contenidos se ha convertido en una necesidad para las plataformas sociales. Pero surgen cuestiones éticas sobre si estos mecanismos de moderación automatizada son una opción viable y respetuosa en el ámbito jurídico.

El principal objetivo es identificar y evaluar las deficiencias que presentan las redes sociales (RRSS) con la moderación de contenidos, así como analizar la labor que desempeña la inteligencia artificial en la moderación y autorregulación. De igual forma, se pretende interpretar y diagnosticar los retos que presentan estos mecanismos. Y: por último, analizar la implantación de sistemas de transparencia y la responsabilidad algorítmica en la moderación y autorregulación de los contenidos de las plataformas sociales con el fin ponderar la seguridad y la libertad de expresión en el entorno digital.

A través de un método de investigación descriptivo se ha podido definir, clasificar y caracterizar el objeto de estudio, con un método deductivo. Se han analizado las políticas y condiciones de uso de las redes sociales Facebook, Twitter (ahora X) e Instagram para verificar su compromiso con la transparencia algorítmica. Se ha realizado una investigación cualitativa a través de la observación investigativa para identificar perfiles de las redes sociales con el fin de identificar problemas realas y fallos en la moderación de contenidos de dichas plataformas, lo

que ha puesto en valor la importancia de contar con una adecuada implementación de la Inteligencia artificial y un especial cuidado con el aprendizaje automático.

Los estudios exploratorios se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado del cual se tienen muchas dudas como ocurre con el tema de la transparencia en la moderación y autorregulación de las plataformas de servicio de red social que es un fenómeno poco conocido (Hernández Sampieri *et al.*, 2014, pp. 90-91). Se trata de una investigación teórica que supone "la actividad sistemática de elaborar, construir, reconstruir, explorar y analizar críticamente los cuerpos conceptuales (esto es, teóricos) en que se enmarcan las distintas áreas del saber" (Barahona Quesada, 2013).

2. LA IRRUPCIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA MODERACIÓN DE LOS CONTENIDOS DE LAS RRSS

La inteligencia artificial es uno de los mayores problemas y paradigmas de nuestro tiempo. Entre sus bondades se configura como una herramienta clave para la innovación, el desarrollo y el avance en múltiples sectores, pero plantea importantes desafíos y cuestiones éticas. Podríamos decir que la IA es el gran paradigma de la sociedad actual. Es unánime la opinión de que la inteligencia artificial es un instrumento clave para la innovación y el progreso, aunque, a su vez, presenta retos y dilemas éticos. Su influencia es tal que abarca prácticamente todos los sectores, desde la automatización de decisiones hasta temas vinculados con los derechos humanos, la privacidad y la seguridad. Su rápida expansión exige una reflexión urgente y profunda para minimizar los potenciales riesgos y garantizar el buen uso social. Su regulación, transparencia y ética en su implementación son factores clave para abordar adecuadamente este desafío.

La irrupción de la inteligencia artificial en la moderación de los contenidos de las redes sociales ha transformado significativamente la manera en que se gestionan las interacciones en línea. A medida que las plataformas buscan equilibrar la libertad de expresión con la necesidad de mantener entornos seguros y respetuosos, la IA se ha convertido en una herramienta clave.

Uno de los fallos que presenta la IA es la escalabilidad, es decir, la capacidad de la IA para procesar grandes volúmenes de datos en tiempo real permite a las plataformas identificar y eliminar contenido inapropiado de manera más rápida y eficiente que los métodos manuales. Esto es especialmente importante en las redes sociales con millones de usuarios, donde el volumen de publicaciones puede ser abrumador.

La IA, según la RAE, es una "disciplina científica que se ocupa de crear programasinformáticos que ejecutan operaciones comparables a las que realiza lamente humana, como el aprendizaje o el razonamiento lógico" (Real Academia Española). La Relatoría Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión de la ONU, va un poco más allá y la define

como "una 'constelación' de procesos y tecnologías que permiten que las computadoras complementen o reemplacen tareas específicas que de otro modo serían ejecutadas por seres humanos, como tomar decisiones y resolver problemas" (Organización de las Naciones Unidas, 2018).

No existe una definición unánime de los sistemas de inteligencia artificial (IA), ya que, por un lado, la IA está en constante evolución debido a los avances tecnológicos, lo que implica que no se trata de un concepto fijo, sino de una tecnología disruptiva que avanza rápidamente. Por otro lado, no hay acuerdo entre los diversos actores involucrados sobre qué constituye exactamente un sistema de IA. No obstante, para poder regular la inteligencia artificial de manera efectiva, es importante alcanzar una comprensión común de qué se entiende por "inteligencia artificial". Esto es fundamental para poder determinar qué aspectos se deben regular y por qué, así como identificar cuáles son los elementos considerados "peligrosos" que requieren supervisión. Cabe destacar que no todos los sistemas de IA representan un riesgo, ya que no todos tienen el potencial de impactar en los derechos fundamentales (Pérez-Úgena, 2024, p. 4). La IA se concibe como la potencialidad de las máquinas para comprender determinados comportamientos de los usuarios y realizar predicciones sobre determinadas acciones.

La inteligencia artificial (IA) es una tecnología que ha experimentado un avance espectacular en poco tiempo, gracias a la combinación de factores como el *big data,* el *blockchain*, la nube, el internet de las cosas, la robótica y la realidad virtual. Aunque la IA no es una invención reciente, ya que sus orígenes se remontan a hace más de 50 años, su impacto actual es enorme y afecta a casi todos los ámbitos de la vida. Una conjunción de factores, como los avances en la potencia informática, la disponibilidad de enormes cantidades de datos y nuevos algoritmos han permitido que se produzcan grandes logros en los sistemas IA en los últimos años (Parlamento Europeo, 2021).

Estas particularidades, implican que la IA cuente con cierta autonomía. Esto se refiere al grado de independencia y autocontrol que poseen los algoritmos al llevar a cabo una tarea sin necesidad de intervención o supervisión humana. Desde la ejecución de instrucciones pre-programadas hasta la capacidad de generar acciones propias basadas en el aprendizaje y la adaptación al entorno, la autonomía de la IA abarca un espectro amplio (Arellano Toledo, 2019; Cotino Hueso, 2022).

2.1. La automatización con propósitos predictivos del comportamiento humano

La proliferación de las plataformas de redes sociales ha incrementado también la complejidad de su moderación, lo que ha generado un reto tanto en términos logísticos como en el ámbito de las relaciones públicas. Hace veinte años, las plataformas de redes sociales necesitaban responder sólo a sus propios usuarios tras incidentes de acoso o trolling (Dibbell, 1993) (Kiesler *et al.*, 2011). Pero la

situación es bien distinta en la actualidad. Para las plataformas de redes sociales ahora dominantes, las técnicas a escala comunitaria se han vuelto cada vez más insostenibles y poco convincentes (Gillespie, 2018).

El volumen y diversidad de contenidos de las RRSS y la rapidez de su publicación y difusión a todos los rincones del mundo ha provocado que los usuarios estén menos vinculados a su círculo social y más por algoritmos de recomendación. Las críticas hacia estas plataformas y su limitada capacidad de autorregulación han aumentado significativamente, especialmente en temas de índole política y social.

Según Boix Palop, la utilización de la automatización con fines predictivos del comportamiento humano juega un papel fundamental, pero también tiene el potencial de generar prácticas injustas o perjudiciales. Uno de los caballos de batalla de estos mecanismos son las censuras por error de muchos contenidos, al no estar capacitados para determinar si realmente ese contenido es o no censurable según las normas de su comunidad. Por ejemplo, en sus normas prohíben la publicación de contenido violento, y si una clínica veterinaria publica un caso clínico de una operación de un animal, ese contenido es censurado por error al identificarlo equivocadamente con contenido violento, cuando es un caso de estudio.

Otro tema cuestionable de los sistemas algorítmicos es la neutralidad. En un principio, los algoritmos gozan de una presunción de neutralidad e imparcialidad basada en el carácter objetivo que suele predicarse en las fórmulas matemáticas. Lo cierto es que desde el momento en el que un algoritmo se programa por una persona con una finalidad concreta, esa neutralidad desaparece. También hay que tener en cuenta que los datos que se seleccionan a través de los algoritmos también pueden tener ciertos sesgos en el momento que aparece el trabajo humano (Pauner Chuli, 2023, p. 120)

Estas prácticas pueden contribuir a la discriminación de personas en condiciones de desigualdad, principalmente aquellos impactados por diferencias estructurales. Además, pueden respaldar decisiones basadas en criterios inapropiados o conducir a predicciones erróneas. Si bien los sistemas automatizados pueden abrir nuevas oportunidades para el acceso, ejercicio y protección de los derechos, también presentan un riesgo significativo de ser utilizados de manera que vulneren esos mismos derechos. Según Boix Palop, los usos de la IA no se limitan a actuar como un instrumento o apoyo en la toma de decisiones, sino que, en realidad, juegan un papel determinante al predeterminar la decisión final de los poderes públicos, delimitando el margen de discrecionalidad a partir de los postulados contenidos en la programación (Boix Palop, 2020, p. 237).

Las RRSS han implementado mecanismos de inteligencia artificial para optimizar su rentabilidad, a la vez que buscan reforzar la seguridad de éstas y minimizar su impacto negativo en la sociedad. La forma de relacionarse entre los usuarios pude ocasionar conflictos que se difunden de manera vertiginosa y pueden afectar, por consiguiente, a la reputación de las personas. Entre los principales desafíos en este entorno digital que ya algunas plataformas están trabajando es en

el anonimato de los usuarios, la divulgación de información personal en espacios públicos y la utilización de la libertad de expresión de manera excesiva.

En cuanto al anonimato, en España ya el presidente del Gobierno, Pedro Sánchez, durante el Foro Económico Mundial de Davos, ha defendido la idea de implantar medidas concretas para poner fin al anonimato en las redes sociales, que trasladará al resto de la Unión Europea. Su medida se centra en vincular los perfiles que se creen con una identidad digital europea (Hernández, 2025).

Este tema de la identidad digital ya lo defendió Sáez de Propios en su tesis doctoral en la que propone que las plataformas de redes sociales implementen mecanismos de autentificación para comprobar que la identidad digital de los usuarios coincida con su identidad real. Igualmente, defiende la protección jurídica de la identidad digital de los usuarios para salvaguardar el respeto a la protección de datos y demás derechos de la personalidad y la reputación online (Sáez de Propios, 2023, p. 565).

Es por este motivo por el que la IA se concibe como una solución potencial para evitar las intromisiones ilegítimas a derechos protegidos de las personas. La gran cantidad de volumen que se genera día a día en las RRSS imposibilita que el trabajo de moderación sea realizado por humanos. Ahí es donde la IA tiene un gran potencial y valor para detectar contenido perjudicial y que pueda atentar contra derechos fundamentales de los usuarios. Este ejercicio debe estar supervisado por humanos para asegurar adecuadas decisiones. Esto ocurre porque los algoritmos no son capaces de detectar el contexto, ni el estilo de comunicación que puede hacer que un contenido esté escrito en un tono metafórico y la IA sea incapaz de comprenderlo.

La IA juega un papel importante en la autorregulación de las RRSS. Ha demostrado una gran eficacia en la moderación de contenido perjudicial o que infringe las normas de uso. Pero su implementación sin la supervisión humana adecuada para proporcionar un contexto o una adecuada interpretación puede hacer peligrar la libertad de expresión de los usuarios al censurar contenido por equivocación. Este tema está siendo motivo de debate internacional, ya que la falta de capacidad de la IA para valorar y contextualizar adecuadamente lo que modera ha provocado que funcione como un instrumento de censura generalizada en lugar de una herramienta de moderación diseñada para salvaguardar la libertad de expresión (Sáez de Propios, 2024, p. 786).

Algunos autores como Larrondo y Grandi han presentado propuestas para perfeccionar la moderación algorítmica en los contenidos de las redes sociales. Su visión empieza con la compatibilización de legislaciones internas de los estados, para respetar los estándares internacionales de libertad de expresión. Igualmente, insisten en el desarrollo de políticas públicas que integren legislaciones protectoras de las condiciones laborales de las personas que trabajan como supervisores humanos, cuya función son las decisiones automatizadas de remoción de contenido. Y, por último, la implantación de transparencia en los términos y condiciones de uso

de las redes sociales, que incluyan auditorías sobre cómo opera la IA en la difusión y eliminación de contenido en línea y, evaluaciones previas de impacto de su IA a los derechos fundamentales (Larrondo & Grandi, 2021, p. 177).

La inteligencia artificial, al aplicar mecanismos de autorregulación, puede llegar a moderar en exceso, y bloquear publicaciones por error y, lo que es más preocupante, restringir la libertad de expresión de los usuarios. Este problema surge cuando la IA identifica un contenido que, desde la perspectiva humana, sería considerado lícito, pero que el algoritmo, al detectar similitudes con contenido prohibido, decide bloquear. Como resultado, las plataformas se ven obligadas a trabajar continuamente en la mejora y ajuste de sus sistemas de moderación (Sáez de Propios, 2023, p. 496).

Esto vislumbra que los sistemas automatizados son aún débiles y necesitan la supervisión humana para poder ser efectivos. En ausencia de una intervención humana que contextualice y traduzca adecuadamente las expresiones en línea, existe un grave riesgo de que las plataformas otorguen prioridad a la inteligencia artificial como moderadora automatizada de contenido en línea. Esto podría resultar en un incumplimiento de los estándares internacionales relacionados con el derecho humano a recibir, investigar y difundir información, ya que la limitación no sería impuesta por una ley necesaria, con un fin legítimo y proporcional, sino por una "constelación" de facto conformada por algoritmos inhumanos (Larrondo & Grandi, 2021, p. 181).

Es importante vincular los mecanismos de autorregulación con la transparencia algorítmica, ya que implica establecer una conexión esencial entre la capacidad de las entidades para regularse a sí mismas y la necesidad de claridad y comprensión en el funcionamiento de los algoritmos. La autorregulación sugiere la capacidad de las partes interesadas para controlar sus propias acciones y comportamientos, mientras que la transparencia algorítmica busca revelar de manera clara y comprensible cómo operan los algoritmos. En conjunto, estos mecanismos buscan lograr un equilibrio entre la autonomía reguladora y la apertura informativa, promoviendo así un entorno ético y responsable en el uso de tecnologías algorítmicas (Sáez de Propios, 2024, p. 787).

Por muy avanzados que sean los sistemas de inteligencia artificial, la realidad es que en la actualidad no se ha resuelto el problema de moderación de contenidos en las RRSS. Lo que sí ha sido una ventaja para estas empresas gestoras de plataformas sociales es la reducción de la carga de trabajo de sus moderadores humanos. Aun así, estas empresas confían en perfeccionar estos sistemas para que puedan identificar contenido ilícito de forma más inmediata y justa que los revisores humanos. Eso les evitaría ver ese contenido ofensivo que hoy en día está causando problemas psicológicos a estos trabajadores.

La quimera para las empresas gestoras de redes sociales es automatizar los procesos, principalmente uno de los más tediosos como es la moderación de contenidos. Los administradores de plataformas sociales justifican el uso de la IA

para poder automatizar la moderación de contenidos a gran escala por la gran cantidad de datos, la imparcialidad de las violaciones y la necesidad de emitir juicios sin pedir que los moderadores humanos los hagan. Se suma la presión de los legisladores que amenazan con imponer obligaciones cada vez más estrictas en la moderación de los contenidos. Y ante esto, los portavoces de las plataformas advierten y justifican que solo la IA puede ser lo suficientemente rápida, precisa y sensible como para cumplir con cualquier obligación mayor. Realmente se ha llegado a un punto en el que solo las soluciones de IA parecen posiblemente viables (Sáez de Propios, 2024, p. 788).

Existen diferentes herramientas de moderación que, en su mayoría, se agrupan bajo el término de "IA". Sin embargo, no todas ellas utilizan realmente la IA avanzada. En algunas plataformas se emplean técnicas de aprendizaje automático para detectar contenido de odio. Estos sistemas aprenden de datos previos para poder identificar patrones y comportamientos problemáticos de manera autónoma (Gorwa, 2020). Realmente la comparación de patrones es un enfoque muy básico que no implicaría realmente un aprendizaje automático que es lo que se denominaría IA.

2.2. Uso de técnicas de aprendizaje automático

La inteligencia artificial se ha utilizado ampliamente para tomar decisiones por los seres humanos. Por ejemplo, ahora se utiliza la IA para detectar información errónea, identificar perfiles problemáticos en línea, identificar los mensajes de odio, personalizar recomendaciones, etc. Aunque estos sistemas de IA están diseñados para tomar decisiones acertadas, la mayoría de los usuarios desconfían de las decisiones tomadas por IA. Entre los factores que llevan al público a dudar de las decisiones de la IA, una razón citada a menudo es la incertidumbre de los usuarios sobre sus procesos de toma de decisiones, es decir, carecen del conocimiento necesario para predecir y explicar las decisiones tomadas por la IA (Berger, 1987).

¿Qué características de la IA pueden inducir incertidumbre y hacer que no sea confiable? ¿Cómo se debe diseñar la IA para que sirva a los objetivos humanos? La creciente delegación de la toma de decisiones a la IA ha creado una urgencia por responder estas preguntas.

El uso de técnicas de aprendizaje automático de la IA moderna puede inducir particularmente incertidumbre entre los usuarios. A diferencia de los sistemas de IA tradicionales que están programados por humanos para seguir reglas creadas por humanos, los sistemas de IA de aprendizaje automático modernos definen y modifican las reglas de toma de decisiones mediante el aprendizaje de los datos. Estas reglas generadas por las máquinas a menudo son desconocidas y pueden ser fundamentalmente diferentes de la lógica humana (Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L., 2016). La investigación hasta ahora sugiere que las personas tienden a tener inteligencia humana simulada y usan su conocimiento sobre

los humanos (es decir, conocimiento antropocéntrico) para reducir la incertidumbre sobre las máquinas; pero este proceso depende de la aplicabilidad del conocimiento antropocéntrico a las máquinas con las que interactúan (Epley, N., Waytz, A., Cacioppo, J. T., 2007).

Para mitigar la incertidumbre de los usuarios, los investigadores en IA explicable (XAI) han estado trabajando para mejorar la transparencia de la IA mediante la construcción de sistemas de IA que puedan explicar sus decisiones a los usuarios (Miller, T., Howe, P. D., & Sonenberg, L, 2017). Aunque algunas investigaciones han encontrado que la transparencia de la IA podría aumentar la satisfacción de los usuarios (Shin & Park, 2019), no está claro hasta qué punto la transparencia de las reglas de toma de decisiones de la IA mitiga la incertidumbre y mejora la confianza y si dicha transparencia puede compensar la pérdida de la agencia humana en la toma de decisiones de la IA de aprendizaje automático. Tampoco está claro cómo los usuarios procesan las explicaciones de la IA.

La falta de transparencia en los sistemas algorítmicos de la IA plantea un desafío a la hora de identificar y abordar los sesgos. Los modelos complejos de aprendizaje automático pueden carecer de interpretabilidad, lo que dificulta discernir cómo se toman las decisiones y si existen sesgos. Un aspecto fundamental en materia de transparencia y acceso a la información es la gestión de los datos utilizados en el entrenamiento de los sistemas de inteligencia artificial, así como los datos de entrada necesarios para su operación, validación y prueba. Este factor resulta elemental, ya que influye directamente en la calidad y solidez del sistema, además de permitir la detección y corrección de posibles sesgos, errores o discriminaciones. Asimismo, cuando la IA procesa datos personales, garantizar la transparencia en su manejo se vuelve una condición indispensable para la protección de este derecho y otros relacionados.

La moderación de contenidos mediante algoritmos de IA plantea varias consideraciones éticas que deben abordarse con cuidado: Transparencia y responsabilidad en las decisiones de moderación de contenido basadas en IA: los usuarios y creadores de contenido deben tener acceso a pautas claras y comprender las reglas y los mecanismos detrás de las decisiones de moderación de contenido tomadas por algoritmos de IA.

El uso de IA plantea importantes preocupaciones éticas que merecen atención. En primer lugar, un consentimiento informado y transparencia en la recopilación y utilización de datos de los usuarios. Las plataformas deben garantizar que los usuarios comprendan los datos recopilados, cómo se utilizan para la manipulación del usuario y el posible impacto en su privacidad y autonomía. En segundo lugar, como garantía de un uso ético de las IA, mejorar la transparencia y la divulgación de la manipulación de los usuarios impulsada por la IA. Las plataformas deben proporcionar explicaciones claras y accesibles de los algoritmos y las técnicas empleadas para la manipulación de los usuarios. La divulgación transparente fomenta la confianza de los usuarios y les permite comprender y

evaluar el impacto de estas prácticas. Así como fomentar la implementación responsable de la IA para promover experiencias positivas de los usuarios: las plataformas de redes sociales deben priorizar el desarrollo y la implementación de algoritmos de IA que mejoren las experiencias de los usuarios, respeten su autonomía y se ajusten a los principios éticos. Las prácticas de IA responsables, incluidas las auditorías periódicas, la mitigación de sesgos y la transparencia algorítmica, pueden ayudar a lograr estos objetivos. Y, en tercer lugar, la necesidad de transparencia y explicabilidad en la detección de desinformación impulsada por IA. Los usuarios deben recibir información transparente sobre cómo los algoritmos de IA detectan y abordan la desinformación. Las plataformas deben esforzarse por explicar la lógica detrás de las decisiones relacionadas con la moderación de contenido y el etiquetado de la desinformación (Rallabandi *et al.*, 2023).

3. MODERACIÓN AUTOMATIZADA DE LOS CONTENIDOS EN LÍNEA

La comunicación en Internet, sobre todo, en las RRSS ha experimentado un crecimiento en los últimos años y se han normalizado las relaciones interpersonales por este canal, así como otras prácticas que tienen que ver con la información, el entretenimiento, la educación, el marketing, los negocios o el networking y las oportunidades laborales. En definitiva, este medio ha transformado la forma en la que los ciudadanos se relacionan entre sí y participan en la sociedad digital. Pero este gran potencial que presentan las redes sociales conlleva, a su vez, riesgos que tienen que ver con la ilimitada intrusión de la tecnología, y con la privacidad.

Que las redes sociales se nutran, principalmente, de contenido generado por sus usuarios en cantidades masivas les ha obligado a aplicar, de forma concomitante, sus propias condiciones de uso y normas comunitarias, así como las leyes aplicables de los países en donde operan, porque el contenido inapropiado puede ser una fuente importante de responsabilidad y afectar, además, a su imagen y reputación. La moderación de contenidos aparece como "una práctica esencial en el ciclo de producción de las redes sociales". Esta actividad se ha convertido en un imperativo en la utilización de las redes sociales ya que protege tanto la propia marca de la red social refuerza la adhesión de los usuarios a las normas de la red, asegura el cumplimiento de las leyes y reglamentos que resulten de aplicación y contribuye positivamente a mantener una audiencia en la red (Roberts, 2017, pp. 33-34).

En este contexto, la IA puede ser útil en la moderación de contenidos que puedan violar los derechos fundamentales, promuevan la violencia o difundan noticias falsas. Sin embargo, su utilización sin la intervención humana adecuada para contextualizar y comprender las expresiones de manera precisa conlleva el riesgo de incurrir en censura previa. Este debate ha cobrado relevancia internacional, puesto que la incapacidad de la IA para contextualizar sus decisiones aumenta la

posibilidad de que se convierta en una herramienta de censura indiscriminada en lugar de una moderación destinada a proteger la libertad de expresión.

La IA suele ser considerada como un conjunto de sistemas tecnológicos automáticos e imparciales tendientes a facilitar la eficacia en la moderación de contenidos en busca de mitigar posibles discursos de odio, discriminatorios, terroristas, etc., y así mejorar la experiencia de sus usuarios y la construcción de ciudadanía (Larrondo & Grandi, 2021, p. 181).

Para optimizar la moderación algorítmica de los contenidos en las redes sociales es esencial llevar a cabo un análisis de la legislación internacional, los informes de organismos globales y las políticas de uso de las plataformas de redes sociales. Las redes sociales han iniciado una carrera por implementar sistemas de autorregulación basados en IA. Estos sistemas deberían considerar la transparencia en sus algoritmos. Las plataformas sociales deben implantar condiciones de uso claras y coherentes, así como adoptar políticas internas de transparencia y modelos de rendición de cuentas sobre el funcionamiento de la IA en la difusión y eliminación de contenidos. Igualmente, es fundamental que realicen evaluaciones de impacto previas en relación con los derechos humanos antes de implementar cambios significativos en sus algoritmos.

La transparencia algorítmica hace que la autorregulación funcione mejor al contar con una previa base jurídica firme, donde los códigos de conducta se conozcan y sean claros y transparentes, según se ha demostrado en un estudio de del Oxford Internet Institute (Barrio Andrés, 2017, p. 138). Tan importante es la transparencia de los algoritmos que, tras la aprobación del RSD, la legislación de la UE sobre los servicios digitales sigue centrándose en mitigar el creciente impacto social de las plataformas en línea. El Reglamento de Servicios Digitales, que entró en vigor el 16 de noviembre, y se aplica a partir del 17 de febrero de 2024, exige una mayor supervisión de los sistemas algorítmicos que constituyen el núcleo del negocio de las plataformas en línea.

La Ley de IA de la UE, en su artículo 14, regula la supervisión humana. Asevera que los sistemas de IA de alto riesgo deben diseñarse de forma que los seres humanos puedan supervisarlos eficazmente. El objetivo de la supervisión humana es prevenir o minimizar los riesgos para los diferentes sectores donde puede interferir la IA como los derechos fundamentales, la salud, la seguridad, la opinión pública, etc. Estas medidas de supervisión deben ceñirse a los riesgos y al contexto de uso del sistema de IA. Estas medidas pueden ser incorporadas al sistema por el proveedor o aplicadas por el usuario. El sistema de IA debe proporcionarse de forma que permita al supervisor comprender sus capacidades y limitaciones, detectar y abordar los problemas, evitar una dependencia excesiva del sistema, interpretar sus resultados, decidir no utilizarlo o detener su funcionamiento. Para determinados sistemas de IA de alto riesgo, cualquier acción o decisión basada en la identificación del sistema debe ser verificada por al menos dos personas competentes.

La Unión Europea (UE), consciente del impacto negativo de la IA en múltiples sectores, ha impulsado iniciativas reguladoras para garantizar su desarrollo y aplicación de manera ética y segura. Su primer paso para unificar sus normativas ha sido la promulgación del Reglamento de Inteligencia Artificial. Esta legislación convierte a la UE en la primera región a nivel mundial en establecer un marco regulatorio completo para los usos de la inteligencia artificial. Se trata de un Reglamento de la Unión Europea y, por lo tanto, es obligatorio en todos sus elementos y directamente aplicable en cada Estado miembro. El pasado 2 de febrero de 2025 entraron en vigor las disposiciones generales y las prohibiciones de prácticas inaceptables como el uso de IA para el reconocimiento emocional en el ámbito laboral. A partir del 2 de agosto de 2025 se han comenzado a aplicar las normas sobre modelos de IA de uso general, gobernanza y sanciones, y a partir del 2 de agosto de 2026, han entrado en vigor las obligaciones para sistemas de IA de alto riesgo. Esta legislación establece las responsabilidades y normas que regirán el uso de una tecnología que, al estar presente de manera continua, está transformando por completo la vida cotidiana de los ciudadanos, al ofrecer tanto oportunidades como riesgos, algunos de los cuales no se han llegado a identificar todavía. Para la Unión Europea, la regulación tiene como objetivo maximizar los potenciales beneficios de la IA como la mejora en los sistemas sanitarios, la seguridad, y sostenibilidad en el transporte, la fabricación eficiente y la generación de energía que garantice unos costes asequibles a la vez que respeta el medio ambiente.

Este Reglamento establece obligaciones para proveedores y usuarios en función del nivel de riesgo de la IA. El avance progresivo de la inteligencia artificial ha puesto de manifiesto la emergencia de múltiples desafíos de índole ética y jurídica. Diversos estudios y análisis han señalado riesgos significativos, entre ellos, la concentración de poder y recursos económicos, la incertidumbre respecto a la evolución y autonomía de estos sistemas, la posibilidad de que se reduzca el control humano sobre la toma de decisiones y el potencial uso indebido de la tecnología por parte de actores individuales o colectivos. Estos factores ponen de manifiesto la necesidad de desarrollar marcos regulatorios y estrategias que garanticen una implementación responsable de la inteligencia artificial, con el objetivo de minimizar sus impactos negativos y promover su aprovechamiento en beneficio de la sociedad.

La Inteligencia artificial es una gran herramienta para las plataformas sociales, ya que tiene un potencial valioso en la rapidez e inmediatez para identificar contenidos violentos, discurso de odio, o cualquier tipo de contenido perjudicial para los usuarios. Pero debe combinarse con la moderación humana que interprete el contenido y contextualice de manera correcta las expresionespara evitar una restricción injustificada de contenidos por error de interpretación (Sáez de Propios, 2024, pp. 791-792).

En cuanto a la supervisión humana, también es necesaria la transparencia, En el caso del artículo 22 RGPD respecto de decisiones IA con tratamientos de

datos personales con IA, las autoridades incluyen la obligación de dar información sobre la "existencia o no de supervisión humana cualificada" y "la referencia a auditorías, especialmente sobre las posibles desviaciones de los resultados de las inferencias, así como la certificación o certificaciones realizadas sobre el sistema de IA". En el caso de sistemas adaptativos o evolutivos, la última auditoría realizada (Agencia Española de Protección de Datos, 2022, p. 24).

Las plataformas de RRSS deben implementar normativas y prácticas responsables en torno al uso de la IA para la moderación de contenidos en línea. En primer lugar, establecer términos y condiciones de uso claros, es decir, reglas comprensibles y bien definidas para los usuarios, en las que se especifique cómo se manejan los contenidos, qué tipo de publicaciones se pueden restringir y/o ocultar, y cómo funciona la moderación automatizada. En segundo lugar, adoptar políticas internas de transparencia y rendición de cuentas. Es importante que las empresas gestoras de RRSS expliquen de forma accesible cómo operan sus algoritmos en la detección, clasificación y eliminación de contenidos. También, deben asumir responsabilidad por los errores y por los sesgos que puedan surgir en sus sistemas con el fin de asegurar que las decisiones automatizadas pueden y deben ser revisadas y corregidas cuando así se requiera. Y, por último, realizar un ejercicio de evaluación y análisis de los efectos potenciales de la IA en derechos fundamentales como la libertad de expresión y la privacidad. Estas auditorías ayudarían a prevenir daños como la censura injustificada o la discriminación algorítmica.

4. TRANSPARENCIA ALGORÍTMICA

La transparencia algorítmica está estrechamente relacionada con la explicabilidad, ya que los resultados y los subprocesos que conducen a ellos deberían aspirar a ser comprensibles y trazables, apropiados al contexto. Los actores de la IA deberían comprometerse a velar por que los algoritmos desarrollados sean explicables. En el caso de las aplicaciones de IA cuyo impacto en el usuario final no es temporal, fácilmente reversible o de bajo riesgo, debería garantizarse que se proporcione una explicación satisfactoria con toda decisión que haya dado lugar a la acción tomada, a fin de que el resultado se considere transparente (UNESCO, 2021).

La Ley de IA [Reglamento (UE) 2024/1689 por la que se establecen normas armonizadas sobre inteligencia artificial] es el primer marco jurídico exhaustivo sobre IA en todo el mundo. El objetivo de las normas es fomentar una IA fiable en Europa. Establece un conjunto claro de normas basadas en el riesgo para los desarrolladores y los implementadores de IA en relación con los usos específicos de la IA. La Ley de IA forma parte de un paquete más amplio de medidas políticas para apoyar el desarrollo de una IA fiable, que también incluye el paquete de innovación en materia de IA, la puesta en marcha de fábricas de IA y el plan coordinado sobre IA. Su objetivo, por tanto, es garantizar la seguridad, los

derechos fundamentales y la IA centrada en el ser humano, y refuerzan la adopción, la inversión y la innovación en IA en toda la UE.

Para facilitar la transición al nuevo marco regulador, la Comisión ha puesto en marcha el Pacto sobre la IA, una iniciativa voluntaria que pretende apoyar la futura aplicación, colaborar con las partes interesadas e invitar a los proveedores e implementadores de IA de Europa y de fuera de ella a cumplir con antelación las obligaciones clave de la Ley de IA (Comisión Europea, 2024).

La Unesco en su Recomendación (n.º 39) señala que se dé "información sobre factores que influyen en una predicción o decisión específica". Esta exigencia es un principio ético de transparencia y comunicación y notificación sobre los datos personales utilizados en el proceso de toma de decisiones y "cómo se llega a los procesos de toma de decisiones automatizadas y de aprendizaje automático" (Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020).

La transparencia puede mejorar la experiencia de usuario en RRSS así como la seguridad de su identidad digital y del contenido que publica. Igualmente, permite a los usuarios tomar decisiones sobre cómo utilizan un sistema algorítmico de toma de decisiones y, de esta forma, juzgar sus posibles consecuencias. Un algoritmo es "una serie finita y discreta de instrucciones que reciben una entrada y producen una salida" (Hogan, 2015). A medida que se delegan cada vez más responsabilidades y procesos importantes a sistemas algorítmicos de toma de decisiones (Diakopoulos, 2017), se presta más atención a la transparencia algorítmica. De ahí que investigadores y responsables de la formulación de políticas en plataformas sociales aboguen por una mayor transparencia como solución para identificar y prevenir los diversos efectos negativos potenciales de estos sistemas.

Hoy en día, las redes sociales deben estar estrechamente vinculadas a la transparencia algorítmica. Para los usuarios de RRSS, la transparencia es esencial al permitirles encontrar información que no es evidente a simple vista y que, en la mayoría de los casos, les resulta difícil entender cómo funciona. Los mecanismos de transparencia ofrecen a los usuarios la posibilidad de acceder a información de un sistema, que permanece oculta. Esto puede influir en su percepción y en la forma en la que interactúan con la tecnología, es decir, conocer las reglas del juego. Una mayor transparencia algorítmica en las redes sociales permite a los usuarios analizar y cuestionar el sistema, aumentar su confianza y poder decidir cómo van a participar en esas plataformas.

Para entenderlo mejor, podríamos comparar la transparencia algorítmica con conocer las reglas de un juego antes de comenzar a jugar. Si los jugadores no conocen éstas, se ven obligados a jugar a ciegas, sin entender por qué ciertas acciones tienen éxito u otras fracasan. Los mecanismos de transparencia actúan como el reglamento de un juego que informa y explica cómo participar, cómo funcionan las estrategias, qué acciones están permitidas y cuáles son las consecuencias de cada una de ellas. Y, además, conocer las reglas permite a los usuarios confiar en el

desarrollo del juego, sin cuestionarlas. Lo mismo ocurriría si en las RRSS no existe la transparencia, donde los usuarios no conocen las reglas del juego, las decisiones pueden resultar arbitrarias y eso provoca un descenso de la confianza.

La transparencia en los sistemas de IA ofrece a los usuarios la oportunidad de conocer aspectos que generalmente están ocultos, lo que puede modificar sus percepciones y creencias sobre cómo funcionan estos sistemas. Al tener más información, los usuarios pueden cuestionarlo y evaluarlo de forma más crítica, y adaptar su comportamiento según el funcionamiento de éstas (Zarsky, 2016).

La opacidad de los algoritmos de IA plantea un reto a la hora de identificar y abordar los sesgos. Los complejos modelos de aprendizaje automático pueden carecer de interpretabilidad, lo que dificulta discernir cómo se toman las decisiones y si existen sesgos (Rallabandi et al., 2023).

La moderación de contenidos mediante algoritmos de inteligencia artificial plantea diversas consideraciones éticas que deben abordarse con cuidado: Las plataformas deben equilibrar el fomento de la libertad de expresión y la mitigación de la difusión de contenidos nocivos, como la incitación al odio, la desinformación y la incitación a la violencia: Los algoritmos de IA pueden perpetuar inadvertidamente los sesgos y la discriminación si se entrenan con datos sesgados o muestran sesgos inherentes en su proceso de toma de decisiones: Los usuarios y creadores de contenidos deben tener acceso a directrices claras y comprender las normas y mecanismos que rigen las decisiones de moderación de contenidos tomadas por algoritmos de IA: La moderación de contenidos que utiliza IA se basa en la recopilación y el análisis de los datos de los usuarios, lo que requiere medidas sólidas de protección de la privacidad y la divulgación transparente de las prácticas de tratamiento de datos.

El uso de la IA en la manipulación del usuario plantea importantes preocupaciones éticas que merecen atención. Las plataformas deben asegurarse de que los usuarios entienden los datos recogidos, cómo se utilizan para la manipulación del usuario y el impacto potencial en su privacidad y autonomía. Las técnicas de manipulación del usuario, incluidas la selección de objetivos emocionales y los desencadenantes psicológicos, plantean cuestiones éticas en relación con el impacto potencial sobre el bienestar del usuario, la salud mental y la manipulación emocional. Los algoritmos de IA se basan en una gran cantidad de datos de los usuarios para crear perfiles y hacer predicciones, lo que exige prácticas transparentes de tratamiento de datos y salvaguardias para proteger la privacidad de los usuarios. Y las técnicas de manipulación de usuarios pueden tener consecuencias no deseadas, como reforzar prejuicios, fomentar la polarización y amplificar contenidos nocivos. Estos daños potenciales ponen de relieve la necesidad de salvaguardias éticas y de un uso responsable de la IA.

Se podrían identificar distintos mecanismos que pueden mejorar la transparencia en la toma de decisiones algorítmicas desde el punto de vista del usuario y de cómo estos comprenden los algoritmos. Uno de ellos sería a través de la

experiencia los usuarios pueden familiarizarse con su funcionamiento según interactúan repetidamente con el sistema, tras la observación de patrones y resultados. Igualmente, los usuarios pueden detectar anomalías y sesgos. Algunos usuarios, al notar posibles fallos o injusticias en los resultados algorítmicos, pueden entender mejor el sistema con el fin de modificar su comportamiento. Y, hay que tener en cuenta, que el conocimiento de cómo funcionan los algoritmos no se distribuye por igual entre todos los usuarios, ya que no todos tienen la misma capacidad o interés para comprender su funcionamiento. Por tanto, aunque los usuarios pueden volverse más conscientes del funcionamiento de los algoritmos a través de la experiencia, esta conciencia no es uniforme ni suficiente para garantizar una transparencia efectiva.

Las auditorías algorítmicas serían otro mecanismo de transparencia clave para mejorar la toma de decisiones basadas en IA, cuyo objetivo es analizar cómo funcionan los algoritmos y qué impactos generan en la sociedad. De esa manera, se pueden identificar sesgos, fallos o consecuencias no previstas en su diseño inicial. Se pueden describir varios enfoques para auditar algoritmos, dependiendo de su visibilidad y responsabilidad (Sandvig et al., 2014). Pero es importante que estas auditorías se lleven a cabo sin la cooperación de los desarrolladores de los sistemas, ya que es probable que incluyan prohibiciones técnicas de recolección de datos o auditorías en sus términos de servicio. Algunas compañías incluso ocultan detalles sobre su funcionamiento para protegerse de la competencia y evitar que los usuarios puedan manipular el sistema en su beneficio. Esto genera un conflicto entre la necesidad de transparencia y la protección de intereses comerciales de las plataformas sociales.

Un tercer tipo de mecanismo para promover una mayor transparencia es proporcionar explicaciones, es decir, proporcionar información sobre cómo y por qué un algoritmo toma una determinada decisión podría ayudar a los usuarios a interpretar mejor su funcionamiento y confiar en sus resultados (Lee et al., 2015).

5. INTENTOS DE REGULACIÓN

Ya en 2017, el Parlamento Europeo emitió una Resolución sobre las implicaciones de los macrodatos en los derechos fundamentales, en el que abordaba temas como privacidad, la protección de datos, la no discriminación, la seguridad y la aplicación de la ley. En esta Resolución, se resalta que el Big Data involucra un proceso automatizado mediante algoritmos informáticos y técnicas avanzadas de tratamiento de datos. Esto incluye el análisis de datos almacenados y transmitidos en tiempo real, con el objetivo de descubrir correlaciones, tendencias y patrones; una práctica conocida como "analítica de macrodatos". En esa Resolución, el Parlamento aborda la cuestión de la transparencia algorítmica, ya que se observa que "algunos casos de utilización de macrodatos involucran la capacitación de

dispositivos de Inteligencia Artificial [IA] como redes neuronales y modelos estadísticos" para predecir comportamientos o situaciones. En consecuencia, se plantea la necesidad de una mayor transparencia en estos procesos algorítmicos. En lo que coincide Wilma Arellano, quien considera que "debe existir la mayor neutralidad posible en el manejo de algoritmos", a pesar de que están desarrollados desde la subjetividad de la persona que ejerce el papel de científico de datos (Arellano, 2019, p. 5).

5.1. Carta de Derechos Digitales en España

La incesante necesidad de implementar mecanismos de autorregulación en el ámbito de las redes sociales ha llevado al Gobierno de España a impulsar el desarrollo de la Carta de Derechos Digitales. Esta medida fue impulsada en 2020, en el marco de los compromisos clave del Plan España Digital 2025.

La Carta de Derechos Digitales no tiene naturaleza jurídica y se concibe como un documento para promover el debate en España y en el ámbito europeo. Ha de entenderse como una agenda normativa y de políticas públicas, a partir de la cual adoptar las medidas necesarias para promover y garantizar los derechos digitales. Cotino Hueso critica la tímida aportación que hace en cuanto a las obligaciones de transparencia que deben satisfacer los protocolos que habrían de implantar las plataformas (Cotino, 2022, pp. 205-215).

Este documento aborda de manera general la transparencia algorítmica como parte de los derechos fundamentales en el ámbito digital. La Carta de Derechos Digitales reconoce la transparencia algorítmica como un aspecto clave para garantizar los derechos de los ciudadanos en la era digital, enfocándose en el derecho a la información clara sobre los algoritmos, su funcionamiento y sus impactos, y asegurando que los usuarios tengan un control adecuado sobre las decisiones automatizadas que los afectan. Aunque la Carta no establece reglas estrictas, sí sienta las bases para un marco normativo más sólido en torno a la transparencia en los algoritmos en el futuro.

5.2. Reglamento de Servicios Digitales 2022/2065

El Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales), es un intento de autorregulación que afecta a los prestadores de servicio de red social como son Facebook, X (antiguo Twitter) e Instagram, objeto de este trabajo. Su aprobación ha supuesto un antes y un después en la regulación del espacio digital y, en especial y en el que nos centramos en esta investigación, en las redes sociales.

Se trata de normas determinantes de la UE para las plataformas en línea. El Reglamento de Servicios Digitales (Digital Services Act) es una propuesta legislativa de la Unión Europea (UE) que busca regular los servicios digitales en línea, incluyendo las plataformas en línea y los intermediarios de Internet. Su objetivo es establecer un marco normativo claro y armonizado para garantizar la seguridad, transparencia y competitividad en el entorno digital. Viene a modificar el marco regulatorio de las redes sociales, especialmente en lo que se refiere a la libertad de expresión, otorgándoles, en la práctica, un ámbito muy amplio de discrecionalidad para poder censurar los contenidos que consideren "inadecuados".

El RSD regula las plataformas en línea y su impacto en los usuarios, y busca aumentar la transparencia y la rendición de cuentas. El reglamento exige que las plataformas, incluidas las redes sociales, expliquen claramente cómo funcionan sus algoritmos y cómo impactan en la difusión de contenido. También incluye medidas para la protección de los menores y la lucha contra el contenido ilegal. Además, exige que las plataformas ofrezcan a los usuarios la posibilidad de desactivar los algoritmos recomendadores, dando mayor control y transparencia a los usuarios.

Por tanto, el Reglamento de Servicios Digitales implica cambios significativos en los servicios de Internet y los derechos de los usuarios. Establece obligaciones para las plataformas en línea, como la transparencia en la forma en que se toman decisiones algorítmicas y la gestión de contenidos ilegales o dañinos. También se requiere que las plataformas en línea implementen medidas para abordar la desinformación y el discurso de odio. Además, el reglamento busca fortalecer los derechos de los usuarios, para asegurar que tengan un mayor control sobre su información personal y garantizar mecanismos efectivos de reclamación y resolución de disputas. También se promueve la interoperabilidad y la portabilidad de datos, lo que permite a los usuarios cambiar fácilmente entre diferentes servicios en línea.

El Reglamento de Servicios Digitales presenta unas obligaciones de transparencia al establecer obligaciones para las RRSS en cuanto a la toma de decisiones algorítmicas y la gestión de contenidos ilegales o dañinos, pues deben proporcionar información clara sobre cómo funcionan sus algoritmos y cómo se selecciona y muestra el contenido a los usuarios. Esto incluye, en particular, saber cómo estas plataformas y motores de búsqueda moderan el contenido y cómo proponen información a sus usuarios; por ejemplo: un servicio de emisión de vídeo en continuo puede utilizar algoritmos para sugerir vídeos que puedan interesar a sus usuarios. Para apoyar la aplicación de estas normas con conocimientos técnicos y científicos de alto nivel, el Centro Común de Investigación (CCI) está creando el Centro Europeo de Transparencia Algorítmica (CETA). Como muchas otras tecnologías digitales emergentes, los sistemas algorítmicos pueden implicar riesgos no deseados, especialmente porque su evolución se produce a un ritmo muy acelerado. Resulta fundamental, por lo tanto, tener una buena comprensión del funcionamiento de la tecnología para supervisar su impacto, convirtiendo el CETA en una herramienta clave para la regulación digital de la Comisión.

5.3. Reglamento de Inteligencia Artificial

En el ámbito de la cadena de valor de la inteligencia artificial (IA), se han establecido obligaciones y responsabilidades específicas con el fin de garantizar una mayor transparencia, particularmente en lo que concierne a la gestión de datos. El Artículo 53.1.a) y b) de la regulación aplicable, junto con los Anexos XI y XII, delinean las responsabilidades de los actores involucrados. Se exige la implementación de medidas de transparencia proporcionales, las cuales incluyen la preparación, actualización y puesta a disposición de la documentación pertinente a los proveedores subsecuentes (Considerando 101 del Reglamento de IA).

Una de las obligaciones clave radica en la transparencia de los datos empleados en el entrenamiento de los modelos, abarcando explícitamente aquellos textos y datos protegidos por derechos de autor. Para abordar la protección de estos derechos, se ha dispuesto la creación de un resumen exhaustivo y accesible al público sobre el contenido utilizado para el entrenamiento de los modelos de IA de uso general. Esta medida, facilitada por la Oficina de IA (Considerando 107 en relación con el Artículo 53.1.d) del Reglamento de IA), busca proporcionar una visión clara del material de entrenamiento. No obstante, es fundamental señalar que la implementación práctica y el alcance de esta obligación aún presentan desafíos en términos de su definición precisa y sus efectos concretos (Ordelin Font, 2025, p. 4).

5.4. Proyecto de ley para la regulación de los algoritmos (Argentina)

En Argentina, se está trabajando en una Ley para la Regulación de los Algoritmos, que busca garantizar la transparencia en el uso de algoritmos, en especial en plataformas de redes sociales. Esta legislación se enfoca en la necesidad de hacer públicos los criterios de los algoritmos, permitiendo que los usuarios comprendan cómo se seleccionan, clasifican o promocionan contenidos.

5.5. Proyecto de "Ley de divulgación de derechos de autor de inteligencia artificial generativa 2024"

El 9 de abril de 2024 se presentó en el Congreso de los Estados Unidos el proyecto de "*Ley de divulgación de derechos de autor de inteligencia artificial generativa 2024*" (*Generative AI Copyright Disclosure Act of 2024*) que tiene por objeto exigir transparencia sobre el material protegido por derecho de autor utilizado para entrenar modelos generativos de inteligencia artificial.

En resumen, aunque no existe una ley única que regule la transparencia algorítmica de las redes sociales de manera global, varios marcos legales internacionales están empezando a abordar este desafío. La tendencia en Europa, especialmente a

través del Reglamento de Servicios Digitales y el Reglamento General de Protección de Datos, está en la vanguardia de la regulación en este ámbito. Sin embargo, la legislación en otros lugares, como Estados Unidos, aún está en fases más tempranas o en discusión, y podría evolucionar hacia regulaciones más estrictas en el futuro.

La Ley de Transparencia de Plataformas Digitales, que aún está en discusión, propone que las plataformas revelen cómo funcionan sus algoritmos, incluyendo la manera en que priorizan, censuran o amplifican ciertos tipos de contenido. También se enfoca en garantizar que los usuarios sean informados de la lógica detrás de las recomendaciones.

6. RESULTADOS

La automatización en la moderación de contenidos presenta un juicio arbitrario que pone de manifiesto un grave problema en la utilización de los algoritmos para decidir qué contenido se permite o se prohíbe en las plataformas de redes sociales, lo que exige contar con sistemas de transparencia. Esta decisión debería depender del juicio humano que pueda interpretar los valores y el contexto en el que se vierte ese contenido a moderar. Además, de que la decisión de un moderador humano solventaría el error de la censura por error que están cometiendo los algoritmos que moderan el contenido en las diferentes plataformas sociales.

Es difícil implementar un sistema de valores que permita a los algoritmos contar con una mayor fiabilidad en la interpretación de los contenidos, ya que los valores humanos son dinámicos, subjetivos y culturalmente variables, mientras que los algoritmos funcionan con reglas fijas y basadas en datos históricos, que, si no cuentan con una actualización permanente, todavía generarían un mayor sesgo y errores interpretativos. En ambas opciones, se necesitaría poner en conocimiento de los usuarios para que se cumpliera el principio de transparencia.

Esto no quiere decir que las herramientas automatizadas no desempeñen ningún papel en la moderación de contenidos, pero no debe operar de forma independiente. Esto significa que se necesite un enfoque combinado donde la IA y la moderación humana trabajen como un tándem para tomar decisiones más justas y precisas. Hasta que la IA no mejore, las empresas de redes sociales se han visto en la obligación de contratar miles de moderadores humanos para llenar ese vacío.

Los responsables ejecutivos de las plataformas sociales se han dado cuenta de que la IA no debería reemplazar por completo el juicio humano. Esta idea sugiere que haya diferentes maneras de integrar la moderación humana y la automatizada para diseñar modelos que aúnan las fortalezas de ambas y mitiguen los errores.

La IA puede funcionar como primer filtro para ejercer una labor de identificación de contenido potencialmente dañino, problemático, o que pueda atentar contra derechos fundamentales, y de esta manera se reduciría la carga del trabajo

de los moderadores humanos. En una segunda fase, la moderación humana podría corregir los errores algorítmicos, como falsos positivos, es decir, contenido legítimo eliminado por error, o falsos negativos, o lo que es lo mismo, contenido que viole las normas de la comunidad que no fue detectado.

La sinergia entre la moderación automatizada y la humana permite un aprendizaje continuo para mejorar los sistemas a través de su entrenamiento. Esa retroalimentación humana ayudaría a reducir los sesgos algorítmicos, y con todo ello se crea un ciclo de aprendizaje para que la IA se vuelva más eficaz.

Otro factor para tener en cuenta es que sería necesaria la utilización de la moderación algorítmica para filtrar el contenido más perjudicial antes de que lo vea un moderador humano. Nos referimos a suicidios, pornografía, violencia, abuso infantil, etc. Esto evitaría que los moderadores tengan que interactuar con este contenido. Esto protegería a los moderadores humanos de enfrentarse a un material traumático. En este caso, los algoritmos identificarían y filtrarían ese contenido lo más rápido posible para que los moderadores humanos pudieran filtrar otro contenido menos perjudicial.

Cada vez más es conocido el lado oscuro de los moderadores de contenidos, considerado el peor trabajo de los últimos tiempos. Estos moderadores son personas encargadas de revisar y filtrar contenido inapropiado, violento o que va en contra de las políticas de las plataformas. A pesar de que su labor es esencial para mantener un entorno seguro en las redes sociales, las condiciones de su trabajo pueden ser extremadamente difíciles y perjudiciales para su salud mental. A continuación, algunos aspectos de esa realidad.

Los moderadores de contenido están expuestos, a diario, a imágenes y videos de violencia explícita, abuso infantil, autolesiones, discurso de odio, entre otros tipos de contenido gráfico y perturbador. Esta exposición repetida puede tener graves consecuencias psicológicas, como el desarrollo de trastorno de estrés postraumático, ansiedad, depresión y otros trastornos mentales.

El testimonio de dos moderadores de contenidos, empleados de Telus International, empresa canadiense subcontratada por Meta —compañía propietaria de Facebook e Instagram— para moderar sus contenidos, refleja las exigencias de esta labor. Su trabajo evita que los usuarios tengan que ver en las pantallas de sus móviles los contenidos más violentos y extremos. A pesar de que firman una cláusula de confidencialidad, sus declaraciones muestran que deben enfrentarse a contenidos como suicidios, violaciones, autolesiones e incluso ataques terroristas en directo, lo que les obliga a reaccionar con rapidez.

Su trabajo sigue una misma instrucción. Cuando hay una denuncia de un usuario, ese contenido les aparece a los moderadores humanos de contenido. En ese momento tienen cuatro opciones: validar, advertir, borrar o escalar. Se borra contenido cuando, por ejemplo, se detecta discurso de odio, *bullying*, e incluso añaden filtros de contenido delicado para una cirugía o imágenes taurinas. Y, el caso más grave, es cuando tienen que escalar ese contenido. Existen diferentes

tipos de escalaciones mediante una escala de jerarquías que puede ser suicidio en vivo, violación, ataque terrorista en vivo, etc. El problema es que los moderadores de contenido tienen un tiempo para tomar una decisión, lo que denominan un *ticket* cada 60 segundos. En declaraciones al programa "Salvados" de la Sexta, un moderador de contenidos aseguraba que "es difícil escalar cosas que luego no van a suceder, por lo que tienes que estar seguro de que se va a suicidar". Ese tiempo de 60 segundos lo tienen que cumplir excepto en los casos de suicidio. Estos casos arruinan sus métricas. Cuando un moderador de contenidos visualiza todo ese contenido, se siente responsable y a la vez una sensación de impotencia de no poder hacer nada real, solo escalarlo (Ramón Lara).

En definitiva, es necesario poner a disposición de los usuarios y mostrar transparencia en todos los procesos y etapas de la moderación de contenido, ya sea por parte de la moderación automatizada de las propias plataformas sociales, como por parte de los moderadores humanos. En este caso, sería necesario mostrar bajo qué criterios, principios y valores actúan.

7. CONCLUSIONES

Este artículo de investigación explora las preocupaciones éticas del uso de algoritmos de IA en las redes sociales. A través de nuestro análisis de la moderación de contenido, la manipulación de los usuarios y la difusión de información errónea, hemos arrojado luz sobre las implicaciones éticas multifacéticas que surgen en este ámbito. Los hallazgos clave de esta investigación destacan la importancia de abordar los sesgos, promover la equidad, garantizar la transparencia y empoderar a los usuarios en el contexto de los algoritmos de IA. La moderación de contenido plantea desafíos para abordar de manera efectiva el contenido dañino y, al mismo tiempo, equilibrar la automatización y el juicio humano. Las tácticas de manipulación de los usuarios y el potencial de desinformación exigen salvaguardas éticas para proteger a los usuarios y mantener la integridad de las plataformas de redes sociales.

Las consideraciones éticas son fundamentales para orientar el desarrollo y la implementación de algoritmos de IA en las redes sociales. Para proteger la privacidad y los datos de los usuarios, las plataformas deben implementar prácticas transparentes de recopilación de datos, almacenamiento seguro y mecanismos sólidos de consentimiento de los usuarios. Garantizar la imparcialidad y mitigar los sesgos en los algoritmos de IA es primordial, lo que requiere perspectivas diversas, auditorías periódicas y un seguimiento continuo. Se debe dar prioridad a la transparencia para generar confianza en los usuarios y fomentar la rendición de cuentas. Además, empoderar a los usuarios con mecanismos de consentimiento, opciones de personalización y herramientas fáciles de usar son esenciales para promover el control de los usuarios sobre sus experiencias impulsadas por la IA.

Las implicaciones éticas de la IA en las redes sociales exigen nuestra atención y esfuerzos concertados. Si priorizamos la privacidad, la equidad, la transparencia y el control de los usuarios, podemos superar los desafíos que plantean los algoritmos de IA y fomentar un entorno ético y responsable para los usuarios de las redes sociales. Las plataformas, los investigadores, los responsables de las políticas y los usuarios deben colaborar para abordar estos desafíos y garantizar que los algoritmos de IA en las redes sociales sirvan a los mejores intereses de las personas y de la sociedad.

El efecto de la inteligencia artificial sobre los derechos fundamentales se refleja en varias áreas, lo que requiere una actualización de las respuestas legales frente a la era digital. Nos encontramos en un momento social en el que las autoridades gubernamentales deben evaluar los peligros de los sistemas de IA y evitar el mal uso de la tecnología para proteger a la sociedad. En el ámbito de las RRSS, los sistemas de inteligencia artificial se utilizan para moderar y autorregular los contenidos de estas plataformas, pero la ausencia de moderadores humanos está provocando una censura previa por error y, por consiguiente, la limitación de la libertad de expresión de los usuarios.

La autorregulación debe implementar sistemas de transparencia en sus procesos algorítmicos para promover la confianza, la equidad y la responsabilidad en estas plataformas. Los algoritmos utilizados por las redes sociales juegan un papel fundamental en la selección y presentación de contenido, al determinar qué publicaciones, noticias o anuncios se muestran a los usuarios. Actualmente, las plataformas mantienen la información de la autorregulación en secreto por razones de propiedad intelectual e industrial, lo que genera desconfianza y viola las expectativas de los usuarios. Se sugiere que se brinde mayor visibilidad sobre los procesos de moderación, lo cual puede requerir cambios en los algoritmos subyacentes para que los resultados sean comprensibles y se ajusten a las definiciones proporcionadas por los usuarios.

Y no sólo la transparencia debe pasar por la moderación de los contenidos, sino por el consentimiento informado de los usuarios, quienes deberían otorgar un permiso explícito sobre el uso de la IA con fines de entrenamiento de sus modelos. Esto permitirá a los usuarios poder decidir si el contenido que publican en sus perfiles sociales puede ser utilizado para entrenar estos sistemas de inteligencia, y proteger así sus derechos de propiedad intelectual y privacidad.

En este sentido, es fundamental que las empresas de redes sociales adopten medidas concretas para mejorar la transparencia de sus algoritmos para conseguir la mayor neutralidad posible. Esto implica proporcionar información clara y comprensible sobre cómo funcionan los algoritmos, qué factores se tienen en cuenta para mostrar u ocultar cierto contenido y cómo se personaliza la experiencia de cada usuario. Debido a la complejidad de los algoritmos y a la falta de transparencia en su funcionamiento interno puede haber ocasiones en las que los algoritmos cometan errores al identificar y moderar contenido, lo que puede llevar a la censura

injustificada de contenido legítimo y a la restricción de la libertad de expresión de los usuarios. Estos errores pueden tener diversas causas, como la falta de comprensión del contexto o la interpretación errónea de ciertos elementos del contenido. Además, los algoritmos utilizados para la moderación pueden estar influenciados por sesgos y subjetividades de los diseñadores, lo que afecta a la imparcialidad y objetividad del proceso. También, los algoritmos enfrentan dificultades para identificar expresiones, especialmente cuando se utilizan palabras clave para ocultar su significado. Hay que tener en cuenta que actualmente la moderación de contenidos combina tanto algoritmos como moderadores humanos para abordar estas limitaciones. Los sesgos inherentes en los datos utilizados para entrenar los algoritmos pueden influir en las decisiones de moderación y limitar la diversidad de opiniones y perspectivas en las plataformas, por lo que la moderación de contenido en las redes sociales plantea desafíos tanto técnicos como humanos.

La transparencia en los algoritmos de las redes sociales también implica revelar cualquier sesgo o discriminación inherente en su funcionamiento. Es importante que las empresas aborden y corrijan cualquier tendencia discriminatoria o injusta que pueda surgir como resultado de los algoritmos, y que informen a los usuarios sobre los pasos que se están tomando para abordar estos problemas. Asimismo, se requiere una mayor transparencia en la forma en que se recopilan, utilizan y almacenan los datos de los usuarios. Los usuarios deben tener claridad sobre qué información se recopila, cómo se utiliza para alimentar los algoritmos y qué opciones tienen para controlar su privacidad y personalización de contenido. Además, sería conveniente informar a los usuarios sobre la posibilidad y los límites de las herramientas de moderación personal.

La labor de los moderadores de contenido es esencial para la seguridad en plataformas digitales, pero conlleva un alto coste emocional. Estos trabajadores enfrentan un flujo constante de material gráfico perturbador, lo que puede derivar en problemas de salud mental. Por ello, es importante que las plataformas de redes sociales reconsideren cómo apoyan a sus moderadores, mejoren sus sistemas algorítmicos para evitarles ver ese contenido e implementen medidas que protejan su salud mental. Esto no solo beneficiaría a los trabajadores, sino que también mejoraría la calidad de su trabajo y la experiencia de los usuarios en las redes sociales. La atención a estas cuestiones es fundamental para equilibrar la seguridad digital y el bienestar tanto de usuarios como de los moderadores.

En el contexto de las redes sociales, comprender la protección jurídica de los derechos fundamentales de la personalidad como pueden ser el honor, la intimidad y la imagen resulta concluyentes para apreciar todas sus complejidades y las implicaciones que conlleva. Es fundamental comprender la naturaleza poliédrica del honor y sus múltiples dimensiones. Las diferentes concepciones del derecho al honor influyen incluso en la autorregulación de los prestadores de servicios, principalmente en los mecanismos de moderación personal de los contenidos,

pues éstos podrán configurar con un mayor o menor grado de sensibilidad y tolerancia según donde el usuario establezca el límite del derecho al honor, que va a estar condicionado por su cultura, sus tradiciones y sus creencias personales. De ahí que sea un aspecto importante por desarrollar en mayor medida por los prestadores de los servicios de redes sociales.

Las redes sociales utilizan algoritmos que pueden influir en la tolerancia hacia ciertas expresiones y contenido. En primer lugar, con el filtrado de contenido donde los algoritmos analizan interacciones de los usuarios (me gusta, comentarios, compartidos) y ajustan lo que muestran en sus *feeds*. Si un usuario interactúa más con contenido de una cierta temática, el algoritmo prioriza ese tipo de publicaciones, lo que puede limitar la exposición a opiniones diversas. En segundo lugar, con la moderación automatizada que pueden censurar o restringir contenido que consideren inapropiado o que infrinja sus normas. Esto puede resultar en la eliminación de expresiones que, aunque no sean ofensivas, son vistas como perjudiciales. En tercer lugar, hay que tener en cuenta que los sistemas de recomendación pueden incentivar a los usuarios a seguir ciertas cuentas o grupos, lo que fomenta la creación de "burbujas" donde se valida solo un tipo de expresión y se ignoran otras perspectivas. En cuarto lugar, se debe tener en cuenta la cultura de cancelación; los algoritmos también pueden amplificar reacciones negativas hacia ciertos contenidos, lo que implicará una mayor censura de esas expresiones. Las consecuencias pueden ser desproporcionadas, lo que puede disuadir a los usuarios de compartir sus opiniones. Y, por último, en cuanto a los ajustes de privacidad y configuración que algunas plataformas permiten a los usuarios ajustar configuraciones sobre qué tipo de contenido desean ver o de qué temas quieren recibir menos información, lo que les da cierta agencia sobre su experiencia. Estos mecanismos pueden afectar la forma en que se perciben y toleran ciertas expresiones, destacando la importancia de la transparencia y la ética en el diseño de algoritmos para fomentar un ambiente de diálogo más abierto y diverso.

Además, la comunicación a través de las redes sociales ofrece infinitas oportunidades de interacción, aunque también presenta riesgos para la privacidad y vulneración de derechos fundamentales. La libertad de expresión puede presumir de haber alcanzado niveles sin precedentes, pero en su contra está que puede ser utilizada de manera perjudicial. En este contexto, la inteligencia artificial surge como una posible solución para la moderación de contenidos, siempre que sea supervisada por humanos para evitar prácticas de censura previa preventiva por error. Las empresas gestoras de redes sociales deben contar con normas de uso y políticas claras, regulaciones para garantizar sistemas de transparencia. La Unión Europea ha tomado la iniciativa con el Reglamento de Inteligencia Artificial. Las plataformas de redes sociales deben desarrollar políticas claras, fomentar la transparencia y la rendición de cuentas en el uso de la IA, y llevar a cabo evaluaciones previas sobre el impacto en los derechos fundamentales para crear un entorno digital ético y justo.

La transparencia y la responsabilidad algorítmica deben ser una obligación para las plataformas de redes sociales ante el impacto creciente de la IA. La opacidad de los algoritmos plantea peligros tanto para la protección de los derechos fundamentales en los usuarios. Esto poner de relieve la necesidad de visibilizar los procesos de diseño y formas de actuar de estos sistemas. La responsabilidad algorítmica exige que los desarrolladores, las empresas y los reguladores tomen la responsabilidad de las repercusiones, especialmente en áreas clave como la toma de decisiones, la privacidad y la justicia. De esta forma, la implementación de políticas claras y responsables, y la transparencia en los sistemas algorítmicos va a permitir proteger los derechos de las personas, además de favorecer un entorno tecnológico más ético y justo. La colaboración entre los diferentes actores es fundamental para enfrentar los retos emergentes y construir un futuro en el que la inteligencia artificial sea utilizada de manera responsable y beneficiosa.

En conclusión, el desarrollo tecnológico, especialmente el aumento del uso de la Inteligencia Artificial (IA) en la toma de decisiones automatizadas, está impactando cada vez más la vida cotidiana de las personas. Esto plantea la necesidad de implementar políticas armonizadas de transparencia algorítmica que permitan a la ciudadanía comprender el funcionamiento de estos sistemas, sus efectos y, al mismo tiempo, identificar y mitigar posibles riesgos. Asimismo, es necesario que los proveedores y desarrolladores de estos sistemas ofrezcan canales de comunicación adecuados para que los usuarios puedan realizar consultas, presentar quejas sobre los resultados obtenidos o informarse sobre los riesgos asociados a su uso.

BIBLIOGRAFÍA

Agencia Española de Protección de Datos. (2022). Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial.Una introducción. https://www.aepd.es/guias/adecuacion-rgpd-ia.pdf

ARELLANO, W. (2019). «El derecho a la transparencia algorítmica en Big Data e Inteligencia Artificial». Revista General de Derecho Administrativo, núm. 50, pp. 1-34. https://www.iustel.com/v2/revistas/detalle_revista.asp?id_noticia=421167

BARRIO ANDRÉS, M. (2017). *Fundamentos del Derecho de Internet*. Centro de Estudios Políticos y Constitucionales.

BERGER, C. R. (1987). Communicating under uncertainty. In M. E. M. Roloff G.R. (Ed.), *Interpersonal processes: New directions in communication research* (pp. 39-62)

BOIX PALOP, A. (2020). Los algoritmos son reglamentos. *Revista De Derecho Público: Teoría y Método, 1*, 223–269. 10.37417/rpd/vol_1_2020_33

Comisión Europea. (2024). *Ley de IA*. https://commission.europa.eu/. Retrieved 13-02-2025, from https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai

COTINO HUESO, L. (2022). Transparencia y explicabilidad de la inteligencia artificial y "compañía" (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta. In L. Cotino

Hueso, & J. Castellanos Claramunt (Eds.), *Transparencia y explicabilidad de la inteligencia artificial.* (pp. 25-70). Tirant lo Blanch.

COTINO, L. (2022). *La Carta de Derechos Digitales*. Editorial Tirant lo Blanch.

DIAKOPOULOS, N. (2017). Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens. In Cerquitelli, T. QUrcia, D., Pasquale, F. (Ed.), *Transparent Data Mining for Big and Small Data* (pp. 25-43). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-54024-5_2

DIBBELL, J. (1993). A rape in cyberspace, or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *The Village Voice, December* (23)

EPLEY, N., WAYTZ, A., CACIOPPO, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 4* (114), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

FJELD, J., ACHTEN, N., HILLIGOSS, H., NAGY, A., SRIKUMAR, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet & Society,* http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

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

GORWA, R. B. R. K., C. (2020). lgorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 3*(1), 1-15.

HERNÁNDEZ SAMPIERI, R., FERNÁNDEZ COLLADO, C., & BAPTISTA LUCIO, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw Hill España.

HERNÁNDEZ, Q. (2025, 22 de enero de). *Sánchez propone acabar con el anonimato en las redes sociales y eliminar todos los perfiles falsos.* La Sexta. Retrieved 04-02-2025, from https://www.lasexta.com/noticias/nacional/sanchez-propone-acabar-anonimato-redes-sociales-eliminar-todos-perfiles-falsos_2025012267910e7747e9a00001e26bc6.html

HOGAN, B. (2015). From invisible algorithms to interactive affordances: Data after the ideology of machine learning. In E. M. Bertino S. A. (Ed.), *oles, Trust, and Reputation in Social Media Knowledge Markets: Theory and Methods* (pp. 103-117). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-05467-4_7

KIESLER, S., KRAUT, R., & RESNICHK, P. (2011). Regulating behavior in online communities. In R. Kraut, & P. Resnick (Eds.), *Building Successful Online Communities: Evidence-Based Social Design* (pp. 77-124). MIT Press.

LARRONDO, M. E., & GRANDI, N. M. (2021). Inteligencia Artificial, algoritmos y libertad de expresión. *Universitas-XXI, Revista De Ciencias Sociales y Humanas,* 177-194.

LEE, M. K., KUSBIT, D., METSKY, E., & DABBISH, L. (2015). Working with machines: The impact of algorithmic and data-driven management on human workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1603-1612)http://dx.doi.org/10.1145/2702123.2702548

Reglamento de Inteligencia Artificial, (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_ES.pdf

MILLER, T., HOWE, P. D., & SONENBERG, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *ArXiv.*

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WACHTER, S., & FLORIDI, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 2* (3) https://doi.org/10.1177/2053951716679679

ORDELIN FONT, J. L. (2025). Copyright and training of generative AI systems: transparency obligations and text and data mining in European regulations.

Organización de las Naciones Unidas, O. (2018). *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión.*

Parlamento Europeo. (2021, 26-03). *¿Qué es la inteligencia artificial y cómo se usa?* https://www.europarl.europa.eu/topics/es/article/20200827STO85804/que-es-la-inteligencia-artificial-y-como-se-usa

PAUNER CHULI, C. (2023). Transparencia algorítmica en los medios de comunicación y las plataformas digitales. *Revista Española De La Transparencia,* (17), 107-136. https://doi.org/10.51915/ret.308

PÉREZ-ÚGENA, M. (2024). *La inteligencia artificial: definición, regulación y riesgos para los derechos fundamentales*. University of Deusto. 10.18543/ed7212024

RALLABANDI, S., KAKODKAR, I. G. S., & AVUKU, O.2023. International Workshop on Intelligent Systems (IWIS), Ulsan, Korea, Republic of, 2023, pp. 1-9, doi: 10.1109/IWIS58789.2023.10284706.

Producciones del Barrio (Producer), & RAMÓN LARA, J. É &. (Director). (14/10/24). *Salvados. Redes sociales: la fábrica del terror.* [Video/DVD]

Real Academia Española. *Diccionario de la lengua española, 23.ª ed.*

ROBERTS, S. T. (2017). Content Moderation. In L. M. Schintler C. (Ed.), *Encyclopedia of Big Data* . Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_44-1

SÁEZ DE PROPIOS, M. (2023). *La frontera entre las libertades informativas y el derecho al honor en las redes sociales. Estudio de Facebook, Twitter e Instagram* (Tesis de doctorado no publicada). https://repositorioinstitucional.ceu.es/entities/publication/72c5a0a4-ef95-4d72-bfd8-d9a2b44e6b48

SÁEZ DE PROPIOS, M. (2024). La IA en la moderación de las RRSS. In A. M. G. Dafonte-Gómez M.I. (Ed.), *Comunicación digital en la era de la Inteligencia Artificial* (pp. 782-798). Dykinson.

SANDVIG, C., HAMILTON, K., KARAHALIOS, K., & LANGBORT, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry.*

SHIN, D., & PARK, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277-284. 10.1016/j.chb.2019.04.019

UNESCO. (2021). *Proyecto de recomendación sobre la ética de la inteligencia artificial.*

ZARSKY, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values, 1* (41), 118-132.