

HURI-AGE

Red Tiempo de los Derechos



Papeles el tiempo de los derechos

ALGORITMOS, REDES SOCIALES Y RACISMO DIGITAL

ALGORITHMS, SOCIAL NETWORKS AND DIGITAL RACISM

Ángeles Solanes Corella

Universitat de València

Palabras clave: algoritmos; redes sociales; racismo digital; esfera pública; discurso del odio.

Keywords: algorithms; social networks; digital racism; public sphere; hate speech.

Número: 24 Año: 2026

ISSN: 1989-8797

Comité Evaluador de los Working Papers “El Tiempo de los Derechos”

María José Añón (Universidad de Valencia)
María del Carmen Barranco (Universidad Carlos
III) María José Bernuz (Universidad de Zaragoza)
Rafael de Asís (Universidad Carlos III)
Eusebio Fernández (Universidad Carlos III)
Andrés García Inda (Universidad de Zaragoza)
Cristina García Pascual (Universidad de Valencia)
Miguel A. Ramiro (Universidad de Alcalá)
María José González Ordovás (Universidad de Zaragoza)
Jesús Ignacio Martínez García (Universidad of
Cantabria) Antonio E Pérez Luño (Universidad de
Sevilla)
Miguel Revenga (Universidad de Cádiz)
Maria Eugenia Rodríguez Palop (Universidad Carlos III)
Eduardo Ruiz Vieytez (Universidad de Deusto)
Jaume Saura (Instituto de Derechos Humanos de Cataluña)

ALGORITMOS, REDES SOCIALES Y RACISMO DIGITAL

ALGORITHMS, SOCIAL NETWORKS AND DIGITAL RACISM

Ángeles Solanes Corella^{1*}

Universitat de València

Resumen: El presente trabajo analiza la relación entre los algoritmos que utilizan las redes sociales y la difusión del racismo digital. Las plataformas tecnológicas, lejos de ser espacios neutros, configuran activamente el debate público mediante sistemas de recomendación diseñados para maximizar la atención del usuario, con el resultado de que los contenidos más polarizadores, entre ellos los de naturaleza racista y xenófoba, encuentran en ese entorno condiciones óptimas para su propagación y normalización. Ante esta realidad, este estudio propone un conjunto de medidas que combinan la regulación jurídica, la responsabilidad de las plataformas y la alfabetización digital de la ciudadanía como ejes de una respuesta que, para ser eficaz, ha de ser necesariamente plural y coordinada.

Abstract: This paper analyses the relationship between the algorithms used by social media platforms and the spread of digital racism. Far from being neutral spaces, technological platforms actively shape public debate through recommendation systems designed to maximise user attention, with the result that the most polarising content, including racist and xenophobic material, finds optimal conditions for its propagation and normalisation. Faced with this reality, this study proposes a set of measures combining legal regulation, platform accountability and digital literacy as the pillars of a response that, to be effective, must necessarily be plural and coordinated.

Palabras clave: algoritmos; redes sociales; racismo digital; esfera pública; discurso del odio.

Keywords: algorithms; social networks; digital racism; public sphere; hate speech.

^{1*} ORCID: <https://orcid.org/0000-0002-4207-6113>. Este artículo ha sido realizado en el marco del Proyecto de Generación de Conocimiento PID2023-147896OB-100 “Derechos Humanos e inmigración: repensar la integración en tiempos oscuros”, financiado por el Ministerio de Ciencia, Innovación y Universidades.

1. Introducción

La consolidación de las redes sociales como espacios centrales de interacción política, cultural y social ha modificado profundamente las condiciones de formación del discurso público. Lejos de constituir meros canales de comunicación, estas plataformas configuran activamente la visibilidad, circulación y jerarquización de los contenidos mediante sistemas algorítmicos cuya lógica responde, fundamentalmente, a intereses económicos. En este contexto, el análisis del racismo contemporáneo exige atender no solo a sus manifestaciones discursivas explícitas, sino también a las infraestructuras tecnológicas que facilitan su reproducción y normalización.

La proliferación de términos como "algoritmo" o "inteligencia artificial" en el debate público no ha ido acompañada, en modo alguno, de una problematización suficiente de sus dimensiones éticas y políticas. Floridi (2024, p. 72) llama la atención sobre un aspecto que con frecuencia se pasa por alto: más que de inteligencia en sentido estricto, lo relevante es la agencia que estos sistemas despliegan, esto es, su capacidad para simular decisiones humanas e incidir así en la organización de la vida social. Ni los algoritmos ni la inteligencia artificial son neutrales; llevan inscritos los sesgos derivados de las condiciones en que fueron diseñados, de los datos que los alimentan y de los objetivos que persiguen. Esta dimensión performativa de las tecnologías digitales ha sido señalada con insistencia por perspectivas críticas que cuestionan la supuesta objetividad de los sistemas computacionales (Noble, 2018; Pasquale, 2016). La ubicuidad de estas herramientas, su secretismo y el poder que acumulan plantean una pregunta incómoda que Harari (2019) ha llevado al extremo: ¿hasta qué punto nos gobiernan más de lo que nosotros las dirigimos?

El presente trabajo parte de la convicción de que los algoritmos de las redes sociales contribuyen de manera decisiva a la difusión y normalización de narrativas racistas, con especial atención a su formulación culturalista contemporánea. Para abordar esta cuestión se adopta una concepción de discurso público inspirada en Habermas (1989), quien lo entiende como el espacio de intercambio de argumentos sobre asuntos de interés común, regido idealmente por principios de racionalidad, respeto y veracidad, tal y como lo ha recordado recientemente Hannon (2023, p. 337). Las condiciones actuales de la esfera pública digital, sin embargo, no solo distan de ese ideal normativo, sino que en muchos aspectos lo contradicen abiertamente.

Este trabajo se estructura en torno a cuatro líneas de análisis. La primera examina la transformación que la digitalización y la plataformización han operado sobre

la esfera pública, poniendo de relieve las tensiones entre las promesas democratizadoras de internet y sus efectos reales sobre el pluralismo deliberativo. La segunda se detiene en la economía de la atención que sustenta el modelo de negocio de las plataformas y en las consecuencias que esa lógica tiene sobre la calidad del debate público. La tercera aborda las formas contemporáneas del racismo y los mecanismos mediante los cuales los sistemas algorítmicos contribuyen a su reproducción y normalización. La cuarta, finalmente, examina el caso español prestando especial atención a la regulación jurídica del discurso del odio, a la incidencia empírica del racismo digital y a algunos episodios paradigmáticos, para desembocar en un conjunto de propuestas orientadas a la construcción de un entorno digital más justo e inclusivo.

La relevancia de este análisis no necesita de una justificación extensa: las plataformas digitales ocupan ya un lugar central en la configuración de la opinión pública y las dinámicas de exclusión racial no solo persisten, sino que se intensifican. El estudio se sitúa deliberadamente en la intersección entre la teoría de la esfera pública, la ética de la inteligencia artificial y los estudios críticos sobre el racismo, apostando por un enfoque interdisciplinar que es imprescindible para comprender los desafíos a que se enfrenta la democracia contemporánea. Habermas (2023) ha denunciado con claridad una regresión política en las democracias occidentales que atribuye, en buena medida, al declive del poder racionalizador de los debates públicos. Preguntarse en qué grado la esfera pública plataformizada contribuye a esa regresión es, precisamente, el punto de partida de este trabajo.

2. La esfera pública en la era de la digitalización

Cuando la web llegó en los últimos años del siglo pasado, muchas voces la acogieron con un optimismo que, visto actualmente, resulta tan comprensible como ingenuo. Todas las personas conectadas parecían estar, por primera vez, en un mismo plano: el anonimato borraba los marcadores visibles de la desigualdad real y el acceso a la información y al conocimiento dejaba de ser, en apariencia, un privilegio de unos pocos. Se defendió esta nueva tecnología como una oportunidad sin precedentes, dejara atrás buena parte de las diferencias y las injusticias estructurales que el mundo físico no había conseguido superar.

Semejante optimismo duró poco. Los prejuicios, la discriminación y la injusticia social están impregnados en nuestras conductas y actitudes, también y a veces con mayor intensidad en el anonimato que protege la identidad digital. De aquel entusiasmo democratizador que justificaba la desregulación de la red se ha pasado a la constatación

de que internet, la web y las redes sociales se han convertido en espacios idóneos para el extremismo, el odio y la causa racista (Bliuc et al., 2018; Klein, 2017).

El ciberespacio ha transformado nuestra forma de concebir el ocio, ha facilitado nuevas maneras de gestionar la economía familiar y ha posibilitado el acceso inmediato a información y recursos culturales de todo tipo. Pero la influencia de la red va mucho más allá de estas dimensiones cotidianas, porque su funcionamiento, moldeado por las grandes empresas tecnológicas que lo sostienen, está diseñado no solo para facilitar su uso sino para fomentar la permanencia del usuario, sirviendo para ello de datos personales que son monitorizados, recopilados y utilizados por complejos sistemas algorítmicos que inducen a actuar en línea según los cálculos más oportunos para el lucro de aquellas compañías, sin que haya una verdadera conciencia de ello por parte de quienes los usan (Zuboff, 2019).

Internet es un espacio descentralizado y desregulado en el que la popularidad prevalece sobre el conocimiento (Siegel, 2008) y cuyo dinamismo hace que cualquier intento de comprensión profunda o regulación quede rápidamente desfasado. La ciudadanía no ha recibido formación para desenvolverse en él más allá de lo que permite la interfaz gráfica de usuario, diseñada precisamente para hacer invisible la complejidad subyacente, y tampoco ha tenido la oportunidad de formarse éticamente para actuar en espacios donde prima el anonimato y la sensación de impunidad. La enorme mayoría de quienes usan estas herramientas desconoce conceptos que afectan directamente a su seguridad y privacidad, lo cual hace de la ciudadanía digital una ciudadanía desarmada frente a poderes que conocen perfectamente sus vulnerabilidades.

2.1. La evolución de la esfera pública: de la publicidad a la plataformización

No puede sorprender que la teoría habermasiana de la esfera pública haya experimentado una profunda relectura a la luz de los procesos de digitalización. Habermas (2023) describe la transformación en curso como una "plataformización" de la esfera pública, un proceso caracterizado por el desplazamiento de los espacios tradicionales de deliberación hacia entornos digitales controlados por actores privados, con todo lo que ello implica para los mecanismos de formación de la opinión pública. Recordemos que Habermas entiende la esfera pública como aquel contexto social en el que puede formarse la opinión pública y cuyo acceso está garantizado para toda la ciudadanía; un espacio sostenido en la conversación entre individuos que debaten sobre asuntos de interés general, al amparo de la libertad de reunión y asociación y de la libertad de expresión de ideas y opiniones (1989). Si los medios impresos, la radio y la

televisión fueron el vehículo del siglo XX, el ciberespacio es hoy esa plaza o cafetería virtual que ofrece oportunidades para la formación de la opinión pública mediante el intercambio de pareceres entre la ciudadanía. Oportunidades que, como se verá, están muy lejos de aprovecharse en condiciones de igualdad.

La esfera pública burguesa que emergió en el siglo XVIII constituyó un espacio crítico en el que los ciudadanos se reunían para contrarrestar el poder de la autoridad política mediante el debate. En los *coffeeshouses* ingleses y los salones franceses prevalecía la autoridad argumentativa por encima de la jerarquía social y se abordaban problemas que hasta entonces no habían sido sometidos al escrutinio público (Habermas, 1986). La imprenta, junto con el capitalismo, fue clave para la creación de naciones al permitir a grupos amplios de personas desconocidas imaginar una comunidad propia gracias al empleo de un lenguaje común (Anderson, 2006), la prensa periódica fue el vehículo que dotaba de conciencia comunitaria a quienes se reunían en esos espacios y que posibilitaba la configuración y propagación de la opinión pública. Ese espacio crítico, sin embargo, no tardaría en ser colonizado.

No tardó la prensa en abandonar esa función crítica para convertirse, como anticipó el propio Habermas (1986, p. 205), en algo "funcional tanto en la manipulación del público como a la legitimación ante él", desplazando la publicidad valorativa y evaluadora por una de índole manipuladora. Con la llegada del ciberespacio y las redes sociales virtuales como nuevos medios comunicativos bajo la tutela de intereses comerciales en un mundo virtual sin regulación, Habermas (2023) advierte la aparición de una comunicación semipública, fragmentada y encerrada en sí misma que distorsiona la imagen de la esfera pública política y amenaza la formación de una voluntad popular medianamente deliberada y basada en el consentimiento real y libre de sus participantes. En la actualidad, la opinión pública es resultado, en buena medida, de una construcción social fruto de una dialéctica que en las redes sociales virtuales está siendo cada vez más secuestrada por las compañías tecnológicas que diseñan los espacios en función de sus intereses comerciales.

La plataformización de la esfera pública se caracteriza por tres elementos (Seeliger y Seignani, 2022): su digitalización, con procesos interconectados mediados por plataformas que prometen una comunicación bidireccional; su mercantilización, con tendencias de instrumentalización y explotación económica de la esfera pública por parte de las propias compañías que diseñan y mantienen las plataformas a través de un modelo de negocio basado en la vigilancia; y su globalización, que ahonda en la

complejidad para comprender el mundo cuando los problemas son de escala transnacional. Asimismo, la transformación de la figura del ciudadano en la del usuario introduce una dimensión adicional de análisis: mientras que la ciudadanía se vincula a derechos y deberes en un marco jurídico-político, la condición de usuario depende de la aceptación de términos contractuales establecidos por plataformas privadas, lo que plantea serias dudas sobre la calidad democrática de estos entornos (Balkin, 2018). En las redes sociales virtuales, la gente puede creer que está contribuyendo a la construcción continua de una opinión pública, pero son solo unas pocas personas y compañías quienes inician o crean tendencias sobre la base de algoritmos diseñados para maximizar su beneficio económico.

2.2. Las Big Tech: soberanas del ciberespacio

En las últimas décadas, un grupo selecto de empresas tecnológicas ha emergido como uno de los agentes más influyentes en la economía global y en la vida cotidiana de las personas, hasta el punto de que hablar de soberanía digital sin hacer referencia a ellas es sencillamente imposible. Las llamadas Big Five (Amazon, Apple, Meta, Alphabet y Microsoft, a las que añadimos X por su relevancia como red social) no solo han liderado la transformación digital, sino que han extendido su influencia de manera determinante a los ámbitos económico, social y político (Vaidhyathan, 2018). No son meras intermediarias para la difusión de información ni simples proveedoras de herramientas para la comunicación, sino que funcionan como gobernadoras de espacios sociales con una burocracia y una estructura efectiva de gobierno que sigue una lógica capitalista centrada en el interés privado de lucro (Belli y Zingales, 2020). Sus usuarios son, en la práctica, trabajadores no remunerados que informan a las plataformas con su comportamiento sobre cómo captar y retener su atención, generando así los datos de conducta que aquellas comercializan.

El poder de las Big Tech radica en unos modelos de negocio diversificados y altamente eficaces en los que el comercio de datos personales es central. Zuboff (2019) bautizó como "capitalismo de vigilancia" esta lógica económica en la que los datos de los cibernautas son la materia prima gratuita que se procesa para modificar conductas. Meta ha transformado el panorama de la publicidad digital construyendo su modelo en torno a la monetización de datos mediante publicidad segmentada de una precisión sin precedentes. Alphabet hace lo propio desde su motor de búsqueda y desde YouTube. La dependencia de los ingresos publicitarios incentiva a estas empresas a diseñar

plataformas adictivas que producen contenido viral y maximizan la extracción de información comportamental. El objetivo es crear efectos de red que generen situaciones de monopolio, así cuanto más gente usa la plataforma, más datos, más valor y más difícil resulta para los competidores entrar (Fernández et al., 2020).

Meta transformó el panorama de las redes sociales y la publicidad digital. Su modelo de negocio se centra en la monetización de datos de usuarios a través de la venta de publicidad segmentada con una segmentación altamente precisa. Alphabet domina el mercado de la búsqueda en internet y la publicidad digital, empleando su motor de búsqueda y otras plataformas como YouTube para recopilar datos de usuarios y ofrecer anuncios dirigidos (Zuboff, 2019). La dependencia en los ingresos de los anuncios dirigidos y la necesidad de extraer más información de los usuarios incentiva a estas empresas a diseñar plataformas adictivas que producen contenido viral y maximizan la información que aquellas pueden exportar del comportamiento de sus usuarios. El crecimiento de estas compañías se basa en el aumento de usuarios, cuanto más grande sea el tamaño de su plataforma, más datos y más valor se puede extraer de ella. Su objetivo es crear efectos de red que resultan en situaciones de monopolio o mercados dominados por unos pocos actores.

Esta lógica de mercado que convierte a los ingenieros en instrumentos del consumismo y el beneficio privado ha sido criticada incluso desde dentro de la propia industria tecnológica. Karp y Zamiska (2025) denuncian que las mentes más brillantes de Silicon Valley, que en otro tiempo colaboraron con los gobiernos para desarrollar tecnologías transformadoras, se dedican hoy a construir aplicaciones para compartir fotografías y algoritmos de marketing, convirtiéndose sin saberlo en portavoces de los intereses de otros. La alianza que impulsó internet y la exploración espacial se ha roto, y con ella la posibilidad de que la tecnología sirva al interés público antes que al lucro privado.

Las Big Tech han acumulado una influencia sustancial sobre la economía y la sociedad, incluso un poder de infraestructura similar al de los Estados (Fernández et al., 2020). Esta situación deja a los Estados soberanos con la necesidad de ejercer su soberanía de forma delegada. Así, si desean censurar contenido, monitorear comportamientos, obtener información personal de usuarios o bloquear información ilegal, carecen en ocasiones de la capacidad suficiente para hacerlo por sí mismos y deben recurrir a las grandes compañías privadas (DeNardis, 2014). El posicionamiento implica necesariamente capital, por lo que aquellos intereses apoyados por personas o

grupos con mayor capacidad económica tienen mayor probabilidad de mantener su estatus de privilegio. El llamado "mercado de las ideas" que los defensores de la libertad de expresión sin límites esgrimen como argumento no es, en la práctica, ni libre ni igualitario. La historia de las plataformas demuestra que las posiciones que triunfan no son necesariamente las más verídicas ni las más útiles para la deliberación democrática, sino las que los algoritmos comerciales priorizan.

2.3. Las redes sociales virtuales: características y funcionamiento

Las redes sociales han alterado profundamente la forma en que se transmiten los mensajes y cómo las personas perciben las interacciones. Facebook lidera el mercado con más de tres mil millones de cuentas, seguido de YouTube con algo más de dos mil millones y medio; Instagram y WhatsApp, ambas de Meta, con alrededor de dos mil millones cada una; TikTok con más de mil millones y medio; y X con 611 millones (Kemp, 2024). Se trata de una ingente cantidad de datos para comerciar por parte de las plataformas, una materia prima gratuita que los gigantes tecnológicos explotan económicamente en su beneficio. Una serie de características definitorias de estas plataformas explican su funcionamiento y su influencia en la transformación de la esfera pública.

En primer lugar, la inmediatez del flujo informativo. La rapidez con que circula la información es una de las principales características de las redes sociales. Esta situación ha traído consigo la conocida como cultura de la inmediatez como una táctica de olvido y un complemento necesario para un neoliberalismo que busca fomentar el consumismo y despolitizar a la ciudadanía. Esta cultura impone un modo de vida consumista que exige que toda necesidad sea satisfecha de inmediato, en un contexto en el que las acciones se convierten en reacciones y en el que los sentimientos se empobrecen, todo lo cual impide aprehender una realidad que solo se alcanza mediante una vida contemplativa (Han, 2023, p. 22). Las manifestaciones y expresiones de odio y racistas se ven beneficiadas por unas herramientas virtuales que animan a sus usuarios a expresar todo aspecto de su experiencia mental sin los filtros que impone la deliberación pausada (Bliuc et al., 2018).

En segundo lugar, la interactividad con una audiencia masiva. Las redes sociales permiten a toda persona registrada comunicarse con cualquier otra usuaria. El diseño de estas plataformas como espacios abiertos y públicos capacita a toda persona conectada a ser lectora o partícipe potencial de todo mensaje emitido en ellas. La posibilidad de

relacionarse con el resto de los usuarios permite a las redes sociales crear un sentimiento de comunidad y de aceptabilidad social de las creencias u opiniones que se emiten online (Farrell y Lockwood, 2023). En tercer lugar, la durabilidad de la información, de tal manera que los mensajes publicados permanecen en el tiempo de forma indefinida y, aunque sean borrados por su autor, existe la posibilidad de que otros usuarios hayan capturado y difundido el contenido. Todo contenido racista o xenófobo que ataque a personas o colectivos puede perdurar ad infinitum.

El anonimato, cuarto lugar, constituye quizás la característica más relevante en cuanto a la repercusión que tiene sobre la comunicación en las plataformas. Las redes sociales permiten construir una identidad digital a través de pseudónimos o avatares, lo que facilita la sensación de impunidad ante los actos llevados a cabo en el ciberespacio y ha hecho del ciberespacio en general y de las redes sociales en particular unas herramientas peligrosas en manos de fanáticos y extremistas y un campo abonado para el racismo (Azani et al., 2020). Un quinto aspecto significativo es la desregulación. Las plataformas autorregulan los entornos virtuales que ellas mismas han establecido y los usuarios deben aceptar sus términos y condiciones sin posibilidad de alteración para acceder a estos espacios públicos virtuales. La gestión de los discursos de odio y racistas es especialmente compleja dado que las sedes de los gigantes tecnológicos se hallan en Estados Unidos, donde la Primera Enmienda constitucional protege la libertad del discurso en formas no permitidas en otros contextos geográficos (Jørgensen, 2020).

Una última característica es esencial: el incivismo. El anonimato y la sensación de impunidad en las redes sociales fomentan un comportamiento incívico, ya que la presión social es prácticamente inexistente en el ámbito virtual. La nueva esfera como un hiperespacio público en el que las personas son animadas a compartir sus respuestas emocionales inmediatas sobre cada evento y ello, unido al enfoque no intervencionista y desregulatorio basado en la libertad del discurso según la doctrina norteamericana, ha habilitado una cultura virtual de lucha, desinhibición y competición por ser popular que ha dejado atrás las normas clásicas conversacionales basadas en la argumentación y el respeto mutuo entre los interlocutores (Clucas, 2020, p. 57). Troleo implica la iniciación deliberada de conflictos y hostilidad entre usuarios con el ánimo de buscar la polémica y la reacción emocional de los destinatarios (Chen et al., 2022), deteriorando la calidad del debate democrático e instrumentalizando la esfera pública virtual al servicio del odio.

2.4. Algoritmos, inteligencia artificial y creación de nichos ideológicos

Los algoritmos son una parte inseparable de las redes sociales y lo mismo puede decirse de la inteligencia artificial, con un uso creciente que influye en muchos de los aspectos vitales del ser humano. Ambas herramientas son empleadas por las plataformas para optimizar la experiencia del usuario con el objetivo de que permanezca más tiempo conectado, exponiéndolo así a los anuncios dirigidos que suponen la principal fuente de ingresos de las redes sociales.

En este contexto, los algoritmos desempeñan un papel central al seleccionar y priorizar contenidos en función de su capacidad para generar interacción. Como señalan Aridor et al. (2023) y Metzler y Garcia (2023), los sistemas de recomendación están diseñados para maximizar el *engagement*, lo que favorece la difusión de contenidos emocionalmente intensos y, a menudo, polarizadores. Mediante el análisis de la información que los usuarios proporcionan a estos sistemas, tales como clics en el botón de "me gusta" o la visualización de vídeos en una categoría específica, estos algoritmos generan recomendaciones y sugerencias que probablemente resulten atractivas para el usuario, creando una espiral que refuerza las preferencias previas y profundiza la radicalización progresiva de las posiciones ideológicas (Russell, 2023).

Uno de los fenómenos más relevantes en este contexto es la aparición de cámaras de eco o burbujas de filtrado, en las que los usuarios son expuestos predominantemente a contenidos que refuerzan sus creencias previas (Barberá, 2020; Menczer y Hills, 2020). Este proceso reduce la exposición a perspectivas divergentes y limita la deliberación racional, favoreciendo la polarización ideológica (Miller et al., 2021). Gonzálvez (2011, pp. 132-133) sostiene que los "nichos digitales" actúan como un obstáculo para el ejercicio pleno de la ciudadanía democrática, ya que, la personalización y el filtrado impuestos por los algoritmos no fomentan una evolución cognitiva, sino que en realidad constituyen una regresión que se aproxima a una forma de "autismo electrónico". Estos nichos limitan severamente las oportunidades para confrontar y contrastar conocimientos diversos y no anticipados, elementos esenciales para el civismo democrático.

Por otra parte, la opacidad de los algoritmos plantea problemas significativos de transparencia y rendición de cuentas. Como advierte Floridi (2024), la falta de comprensión sobre el funcionamiento de estos sistemas debilita la autonomía de los usuarios y dificulta la identificación de sesgos. Los sistemas algorítmicos han sido descritos metafóricamente como una "caja negra" (Pasquale, 2015), una metáfora que sugiere la ausencia de transparencia respecto de su funcionamiento interno. No es

posible discernir de manera clara lo que sucede dentro de ellos, dificultando así cualquier intento de supervisión o intervención. Esta problemática se ve exacerbada por la carencia de diversidad en el personal de las Big Tech. Sax et al. (2020) han observado que Facebook, consciente de cómo sus algoritmos explotan la atracción mental por la conflictividad y temática polarizadora, no toma medidas para remediar esta situación debido a su rentabilidad, un modelo de negocio que contribuye a dañar tanto a la sociedad como a la democracia misma. Laub (2019) afirma que los algoritmos y la inteligencia artificial, con tal de fomentar la participación e interactividad de los usuarios en las redes sociales, están facilitando de forma accidental la propagación de contenido extremista, engañoso y falso, agravando los problemas creados por la desinformación masiva (Solanes y Hernández, 2025).

Además, como señala Susskind (2025), no puede obviarse que junto a las redes la inteligencia artificial constituye el desafío definitorio de nuestra época, y afrontarlo exige ante todo aprender a pensar sobre ella con rigor y sin dejarse arrastrar ni por el entusiasmo irracional ni por el pánico. Al igual que ocurre con las redes, en su opinión, pueden identificar entre los riesgos más relevantes de los sistemas de IA el sesgo inaceptable y los fallos de rendimiento, que constituyen una categoría específica de amenaza distinta de los riesgos existenciales o catastrofistas que dominan el debate público. A ello añade el riesgo derivado de la dependencia en sistemas que operan de manera poco comprensible o directamente inexplicable para quienes los usan y para quienes deberían regularlos. Nos encontramos así ante una opacidad estructural que hace de la rendición de cuentas una tarea extraordinariamente difícil y que convierte en urgente la necesidad de nuevos marcos regulatorios y procesos democráticos capaces de gobernar una tecnología cuya evolución supera con creces la capacidad de respuesta de las instituciones políticas y jurídicas existentes.

3. Racismo en la gobernanza algorítmica

El racismo contemporáneo ha pasado de apoyarse en supuestas diferencias naturales entre razas a articularse en torno a la idea de incompatibilidad cultural (Aguilar y Buraschi, 2019), Dicha mutación que no implica su extinción sino su adaptación a un contexto social en el que las justificaciones biologicistas resultan ya socialmente inaceptables. Wistrich (1999, p. 2) ya lo describió con precisión al entender que el racismo sigue operando como una forma de heterofobia que transforma diferencias reales o imaginadas en categorías rígidas que legitiman la exclusión. La raza no existe como

realidad biológica y las teorías de supremacismo racial son científicamente falsas, moralmente condenables, socialmente injustas y políticamente peligrosas; sin embargo, el racismo persiste en la sociedad adoptando formas nuevas y más difíciles de identificar.

El racismo puede entenderse como un sistema de dominación que se manifiesta en múltiples dimensiones, interpersonal, institucional y cultural, y que se basa en la racialización de diferencias (Aguilar y Buraschi, 2016). Este proceso implica la construcción de categorías que se presentan como naturales, aunque en realidad son producto de dinámicas históricas y sociales. Barker (1981) denominó "nuevo racismo" a esta modalidad que, en lugar de fundamentarse en la biología, apela a la diferencia cultural como argumento para la exclusión, confiriéndole una apariencia de legitimidad que el racismo biologicista no tiene en el contexto político actual. En las sociedades contemporáneas, el racismo se articula frecuentemente a través del etnocentrismo, la xenofobia y el identitarismo. Como destacan Bizumic y Duckitt (2012), el etnocentrismo implica la centralidad del grupo propio y la desvalorización de los otros. Esta lógica se traduce en discursos que defienden la preservación de una supuesta identidad cultural homogénea, especialmente en el contexto europeo (Bratt, 2022).

Asimismo, la xenofobia se presenta a menudo como una reacción natural frente a la diferencia, aunque investigaciones recientes cuestionan esta interpretación (Saito, 2021). En realidad, se trata de una construcción social que legitima políticas de exclusión bajo argumentos de seguridad o cohesión social. El racismo culturalista se integra en discursos aparentemente democráticos, lo que facilita su normalización. Las ideologías extremistas presentes en el ciberespacio lanzan opiniones contrarias a las personas migrantes o a minorías y colectivos racializados que se están convirtiendo cada día más accesibles y generalizadas (Farrell y Lockwood, 2023). En un escenario de victimismo y actitudes defensivas de posiciones de extrema derecha, estas ideas se normalizan como inocuas y se introducen subrepticamente en debates generales sobre globalización y desplazamiento humano. Teorías como el gran reemplazo, el genocidio blanco o el aceleracionismo han encontrado en las redes sociales una identidad colectiva transnacional que hace difícil su erradicación (Bliuc et al., 2018).

Hay que advertir que en todas las maneras en las que el racismo se expresa subyace la idea de desigualdad de poder, que da paso a la discriminación y a la concentración excluyente de privilegios. Es racista quien, a pesar de la inexistencia de la raza como concepto aplicable a la especie humana, toma en consideración la categorización biológica y racializada del ser humano, entendiendo que cada persona

pertenece a una tipología concreta, cada una de las cuales es esencialmente diferente, siendo superior aquella de la que el racista forma parte. Aunque la manifestación extrema de estas actitudes es minoritaria, el racismo está en auge y aumentan los grupos y personas que difunden odio racial a través de las vías que ofrece el ciberespacio (FRA, 2024). Las plataformas de redes sociales se han erigido como un campo de guerra para la lucha simbólica, comunicativa e ideológica (Fuchs, 2019), pero es el odio quien está ganando la batalla. Ganesh (2018) describe una cultura digital del odio que, de manera paradójica, ha sabido explotar los canales que prometían servir de vehículo a colectivos marginados para que su voz fuera escuchada.

3.1. Sesgos algorítmicos y reproducción de desigualdades

Que los algoritmos pueden contribuir activamente a la reproducción del racismo no es ya una hipótesis, es una constatación documentada. Basta con reproducir el pasado para generar discriminación en el presente, aunque nadie lo haya pretendido. Los sistemas de reconocimiento facial con mayores tasas de error en personas racializadas son el ejemplo más documentado (Perkowitz, 2021), pero están lejos de ser el único. La complejidad de los sistemas de aprendizaje profundo y aprendizaje automático agrava el problema, ya que, cuando la razón detrás de una decisión no puede ser explicada ni por quienes programaron el sistema, refutar una acusación de sesgo se convierte en un desafío casi insuperable.

Noble denomina "algoritmos de la opresión" a aquellos que reproducen online la discriminación a la que están expuestos determinados grupos. Esta autora anticipa que la inteligencia artificial será uno de los mayores retos en términos de derechos humanos en el futuro y sostiene que el poder de los algoritmos intensifica las relaciones sociales opresivas contra colectivos racializados. Ofrece numerosos ejemplos de cómo el algoritmo de Google ha actuado de manera racista y discriminatoria: etiquetaba a personas afroamericanas en los resultados de búsqueda de imágenes con términos ofensivos; presentaba la Casa Blanca como resultado de la búsqueda de un término ofensivo contra la comunidad afroamericana; y proponía resultados de búsqueda para "women cannot" y "women should" que incluían estereotipos de género como "women should be slaves" (Noble, 2018, pp. 1, 6, 15, 20-22 y 159). Google Photos ha tenido también problemas similares al etiquetar a personas racializadas de forma automática como gorilas por utilizar una base de datos que no empleaba una muestra adecuada de personas no caucásicas (Lea, 2020).

Silva y Kenney (2018) presentan una taxonomía detallada de los diferentes tipos de sesgos algorítmicos: el sesgo asociado a la información de entrenamiento; el sesgo del enfoque algorítmico, producido por la inclusión o exclusión de información específica; el sesgo en el procesamiento algorítmico, referido a los prejuicios incorporados en la programación del propio sistema; el sesgo del contexto de transferencia, que ocurre cuando la información se aplica a un contexto inapropiado; el sesgo de interpretación, que surge cuando los usuarios interpretan resultados ambiguos a través de sus propias actitudes parcializadas; el sesgo de falta de transparencia en los resultados; el sesgo del consumidor en las plataformas digitales; el sesgo del bucle de retroalimentación, que muestra cómo los algoritmos aprenden de las interacciones del usuario y ajustan resultados en función de comportamientos anteriores, reforzando y perpetuando sus prejuicios; y el sesgo de automatización, que se presenta cuando los resultados generados son percibidos como hechos objetivos en lugar de predicciones con un determinado nivel de confianza.

El popular bot conversacional Tay que Microsoft creó para X constituye un ejemplo paradigmático de sesgo algorítmico derivado del aprendizaje. Este experimento fue cancelado por la empresa dieciséis horas después de su lanzamiento porque Tay se había convertido en una cuenta que emitía contenido racista de forma continua, al haberse alimentado de las interacciones sesgadas de sus usuarios. También la aplicación FaceApp, en su modo belleza, creaba una piel más clara, unos ojos más redondeados y narices más pequeñas a las personas que querían editar una foto revelando así los sesgos eurocéntricos incorporados en sus algoritmos de mejora estética. La doctrina especializada sobre sesgo y equidad en el aprendizaje automático permite identificar una tipología de prejuicios algorítmicos que resulta especialmente útil para comprender cómo los sistemas de inteligencia artificial reproducen y amplifican discriminaciones preexistentes.

Mehrabi et al. (2021) distinguen, entre otras categorías, los sesgos derivados de los datos de entrenamiento, que reflejan las desigualdades históricas incorporadas en la información con que se alimenta el sistema; los sesgos propios del algoritmo, vinculados a las decisiones de diseño y a los criterios de optimización adoptados por sus programadores, quienes inevitablemente proyectan en el código sus propias experiencias, valores y puntos ciegos; y los sesgos emergentes del propio proceso de aprendizaje automático, en el que la información tendenciosa con que se entrena el sistema no solo se reproduce sino que se amplifica en sucesivas iteraciones, generando

resultados cada vez más alejados de la equidad. La peligrosidad específica de esta última categoría reside en que los sistemas de aprendizaje profundo aprenden de sus propias predicciones anteriores, lo que convierte un sesgo inicial en una distorsión estructural que se consolida y profundiza con el tiempo sin que haya ningún mecanismo automático que lo corrija.

3.2. Redes sociales y normalización del racismo

Las redes sociales no han creado el racismo, pero sí fomentan las condiciones más favorables que este ha tenido jamás para su difusión y normalización. La combinación de anonimato, alcance masivo e inmediatez facilita la expresión de discursos extremos sin las consecuencias sociales que acarrearían en el mundo físico (Kim, 2020). La lógica de la viralidad hace el resto: los contenidos polémicos y emocionalmente intensos se propagan con una velocidad y un alcance que ningún otro medio había permitido. El racismo culturalista se adapta extraordinariamente bien a este entorno porque puede presentarse como discurso político legítimo, como opinión discrepante o incluso como defensa razonada de la identidad. La repetición constante de esos mensajes, además, contribuye a generar una percepción de legitimidad basada en la mera frecuencia de exposición: lo que se ve y se escucha mucho acaba pareciendo normal y verdadero.

Las cámaras de eco intensifican este proceso al aislar a los usuarios en entornos ideológicamente homogéneos. En estos espacios, las narrativas racistas se refuerzan mutuamente y evolucionan hacia formas más radicales. Clucas (2020) describe la existencia de una subcultura de debate en los medios sociales alejada de la estructura tradicional del mundo físico. En este contexto, los usuarios se autorregulan sobre un consenso implícito en torno a qué se puede decir o no, y emplean el humor y el troleo como estrategias para expresar lo que desean, forzando al receptor a aceptar el contenido por estar amparado en el humor o el ingenio, lo cual facilita la normalización del extremismo y del racismo en el debate. El humor, si bien puede parecer inofensivo, es en realidad un vector muy eficiente en la difusión y generalización del racismo (Weaver, 2011).

La moderación de contenidos presenta limitaciones significativas. La falta de recursos, los problemas de contextualización y la dependencia de sistemas automatizados dificultan la detección de formas sutiles de racismo (Gongane et al., 2022). El trabajo del OBERAXE ilustra estas limitaciones: en 2023 notificó a cinco

plataformas un total de 2.655 contenidos de discurso del odio, de los cuales se retiró solo el 49,4%, con TikTok a la cabeza (77,7%) y X en última posición (31,5%). El 61,99% del contenido retirado fue puesto en conocimiento de las plataformas por los llamados *trusted flaggers*; sin embargo, ni siquiera ante la denuncia de estas personas que la propia plataforma considera expertas se alcanza la mitad de contenido eliminado de las redes sociales (OBERAXE, 2024).

La interpretación absolutista de la libertad de expresión en ciertos contextos legales favorece la persistencia del discurso racista. Como señala Waldron (2012), esta concepción ignora el impacto del discurso de odio sobre la dignidad de las personas y contribuye a su normalización. Por su propia virtualidad y falta de contacto o visualización física del interlocutor, las redes sociales facilitan la desconexión moral de la agencia humana respecto al daño que pueden causar las acciones de quien publica contenido perjudicial para otras personas. Bandura (2002) estableció varios mecanismos que facilitan esa disociación: la justificación del acto, el uso de eufemismos para hacer referencia a la conducta, la realización de comparaciones ventajosas, el desplazamiento o difusión de la responsabilidad, la indiferencia o distorsión de las consecuencias y la deshumanización o culpabilización a la víctima.

A menudo, las personas que comparten en redes sociales sus experiencias con el racismo son silenciadas tanto por humanos como por los propios sistemas de moderación de las máquinas, lo que implica una doble victimización. Los algoritmos de moderación no solo fallan al detectar el discurso de odio, sino que activamente silencian a las víctimas que intentan visibilizarlo.

4. Racismo en redes sociales: especial atención al caso español

En España, el ordenamiento jurídico reconoce la libertad de expresión como un derecho fundamental, pero no con e carácter absoluto. La propia Constitución establece sus límites y la legislación y la jurisprudencia los han desarrollado específicamente en el ámbito del discurso del odio. Esta posición contrasta con la doctrina norteamericana, donde autores como Baker (2009) afirman que la prohibición de mensajes racistas fomenta un victimismo en quienes los emiten que les hace sentirse perseguidos por el sistema. Sin embargo, también hay quienes emplean ese mismo argumento de la democracia para llegar a la conclusión opuesta y entienden que el discurso del odio tiene un grave efecto de silenciamiento sobre las víctimas, quienes dejan de actuar en la comunidad, se repliegan sobre sí mismas y pierden la confianza y la autoestima

colectiva necesarias para la participación en igualdad (Alonso, 2017).

El artículo 510 del Código Penal, introducido en 1995 y reformado principalmente en 2015 mediante la Ley Orgánica 1/2015 y completado por la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, constituye la principal herramienta jurídica para combatir los delitos de odio y el discurso del odio en España. La reforma de 2015 obedecía a dos motivos: ajustar el ordenamiento a la interpretación que ofreció el Tribunal Constitucional en su sentencia 235/2007 con respecto al delito de negación del genocidio, y adaptarse a la Decisión Marco 2008/913/JAI del Consejo. En su redacción vigente, el primer párrafo del artículo 510 castiga con pena de prisión de uno a cuatro años y multa de seis a doce meses a quien públicamente fomente, promueva o incite, ya sea directa o indirectamente, al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo o contra una persona por su pertenencia a aquel por motivos racistas, antisemitas, antigitanos u otros referentes a la ideología, religión o creencias, situación familiar, etnia, raza, origen nacional, sexo, orientación o identidad sexual, por razones de género, aporofobia, enfermedad o discapacidad.

La pena se agrava cuando los hechos se realizan mediante el uso de medios de comunicación social, internet o tecnologías de la información que hagan la conducta accesible a un elevado número de personas (artículo 510.3). El Tribunal Supremo ha afirmado que las redes sociales "intensifica[n] de forma exponencial el daño de afirmaciones o mensajes que, en otro momento, podrían haber limitado sus perniciosos efectos a un reducido y seleccionado grupo de destinatarios", ya que "quien hoy incita a la violencia en una red social sabe que su mensaje se incorpora a las redes telemáticas con vocación de perpetuidad" (STS 4/2017, FJ 2.º). Este mismo Tribunal Supremo determinado también que la ofensa que se halla dentro del discurso del odio es en sí misma la conducta delictiva, ya que "son expresiones que, por su gravedad, por herir los sentimientos comunes a la ciudadanía, se integran en la tipicidad" (STS 72/2018, FJ único).

La regulación del discurso del odio en España no está exenta de críticas doctrinales. Como señala Tapia (2021, p. 284) la amplitud de la redacción puede abarcar prácticamente todo, como aquellas conductas cercanas al contexto artístico, la crítica política o la opinión contestataria. De esta forma, se ha creado un totum revolutum de conductas o manifestaciones que tienen como denominador común la concurrencia de un supuesto sentimiento de odio en el autor hacia la víctima. En todo

caso, la redacción del precepto solventa dudas importantes en torno a la legitimidad activa e incorpora la incitación indirecta, acercándose a la línea del derecho internacional e incluso la jurisprudencia del TEDH (Gascón, 2016).

4.1. La realidad empírica del discurso del odio en España

Las estadísticas oficiales sobre los delitos de odio evidencian que, tras el descenso puntual registrado en 2024, estos actos punibles han retomado su tendencia al alza. Entre los factores que ayudan a explicar este incremento destacan el fuerte aumento de la islamofobia y la facilidad para su comisión que ofrecen las nuevas tecnologías. El total de hechos conocidos registrados por las Fuerzas y Cuerpos de Seguridad ascendió en 2025 a 2.417, un 23,63% más que en 2024 (1.955) y por encima incluso de los 2.268 de 2023, de modo que se recupera la senda creciente de años anteriores. De esa cifra, 2.242 fueron delitos y 175 infracciones administrativas y otros incidentes, partida esta última que se duplicó con creces respecto al ejercicio previo (+103,49%). Por ámbitos, los delitos de motivación racista o xenófoba supusieron el 41,66% del total de delitos (934 hechos), muy por encima de los segundos en la lista, los dirigidos a las víctimas por su orientación sexual e identidad de género (25,47%; 571 hechos), y de los cometidos por motivos de ideología (10,75%; 241 hechos). En términos de variación interanual, los crecimientos más acusados se dieron en la islamofobia (+133,33%), la disofobia (+90,32%) y el antisemitismo (+86,49%), mientras que el único ámbito que retrocedió fue la discriminación generacional (-31,25%) (Ministerio del Interior, 2025).

Atendiendo a la naturaleza de los hechos, la tipología más común fue la amenaza (446 hechos; 18,45%), seguida muy de cerca por las lesiones (441; 18,25%) y por la promoción o incitación pública al odio (319; 13,20%), a las que siguen los daños, las injurias y el trato degradante. En cuanto al lugar de comisión, el escenario más habitual fueron las vías de comunicación (877 hechos; 36,28%), seguidas de las viviendas (494; 20,44%), las instalaciones (476; 19,69%) y los establecimientos (371; 15,35%). La distribución mensual concentró el mayor número de hechos en julio (259), octubre (250) y septiembre (221), alterando la estacionalidad del año anterior. La tasa de esclarecimiento se situó en el 65,66%, con 1.585 hechos esclarecidos y 1.018 personas detenidas o investigadas, y el fenómeno presentó una marcada desigualdad territorial: frente a una tasa nacional de 4,92 hechos por cada 100.000 habitantes, destacaron muy por encima la Ciudad Autónoma de Melilla (21,89), la Comunidad Foral de Navarra

(15,65) y la Ciudad Autónoma de Ceuta (10,77) (Ministerio del Interior, 2025).

El perfil de las personas afectadas y de los presuntos responsables completa el cuadro. En 2025 se contabilizaron 2.294 victimizaciones, un 16% más que el año anterior, con un claro predominio de las víctimas de sexo masculino (62,34%) y una presencia destacada de menores de edad (12,95%); el grupo de edad más victimizado fue el de 26 a 40 años (33,22%). La mayoría de las víctimas eran de nacionalidad española (60,46%), siendo las de origen marroquí las más numerosas entre las extranjeras (9,98%), por delante de las colombianas (4,32%) y venezolanas (2,22%). En cuanto a la autoría, el responsable detenido o investigado fue mayoritariamente varón (78,49%, frente al 81,88% de 2024), de nacionalidad española (75,15%) y situado de nuevo en la franja de 26 a 40 años (24,26%). Por último, el componente digital del fenómeno resultó especialmente significativo: los delitos e incidentes de odio cometidos a través de internet y redes sociales alcanzaron los 283 hechos conocidos (+31,63%), con incrementos extraordinarios en islamofobia (+450%) y antisemitismo (+109,09%), siendo internet (36%) y las redes sociales (32%) los medios más empleados, seguidos de la telefonía y las comunicaciones (20%) (Ministerio del Interior, 2025).

Sin embargo, estas cifras están muy por debajo de la realidad dado los elevados niveles de infradenuncia existentes en este contexto. Los motivos para no denunciar son múltiples: la falta de confianza en el sistema, la inutilidad percibida de la denuncia, el miedo a la policía, el desconocimiento de la normativa y de los procedimientos o el temor a una posible expulsión en caso de personas migrantes. De hecho, muy pocos de los incidentes que encajan con la tipología delictiva del delito de odio son realmente denunciados a alguna organización o a las autoridades, incluso las agresiones físicas, la mayoría de las cuales no son puestas en conocimiento de estas últimas (FRA, 2021).

La fuente más sistemática para dimensionar el discurso de odio en el entorno digital español es el sistema de monitorización del Observatorio Español del Racismo y la Xenofobia, cuyos resultados se recogen anualmente. A lo largo de 2024 se notificaron 2.870 contenidos considerados de odio racista, xenófobo, islamófobo, antisemita o antigitano a las cinco plataformas monitorizadas (Facebook, X, Instagram, TikTok y YouTube), susceptibles de ser constitutivos de delito, de infracción administrativa o de vulnerar las propias normas de uso de los servicios. El reparto de notificaciones estuvo encabezado por X (758 casos; 26%) y Facebook (727; 25%), seguidos de Instagram (538; 19%), TikTok (478; 17%) y YouTube (369; 13%).

El dato más revelador para un análisis de la moderación algorítmica es el bajo

nivel de respuesta de las plataformas: del total notificado solo se retiró el 35% de los contenidos (1.010 casos), y la tasa de retirada descendió catorce puntos respecto a 2023. Esta cifra global esconde, además, una diferencia decisiva según el canal de notificación. Cuando el aviso se cursó desde un perfil de usuario corriente, únicamente se eliminó el 9% de los contenidos (272 casos), mientras que la retirada ascendió al 26% (738) cuando se notificó mediante la figura del *trusted flagger* o comunicante fiable. En otras palabras, los mismos contenidos reciben un tratamiento desigual en función de quién los reporte, lo que evidencia que los circuitos automatizados de moderación dirigidos a la ciudadanía resultan notablemente ineficaces: solo el 4% de los contenidos se retiró en las primeras 24 horas, el periodo más relevante para contener la viralización. El 65% de las notificaciones (1.860 casos) no fue retirado en absoluto, pese a que el 96% de ellas infringía las propias normas de conducta de las plataformas. Esta brecha entre la norma declarada y la actuación efectiva de los sistemas de moderación constituye, en sí misma, una forma de discriminación mediada por la arquitectura técnica de las redes.

En cuanto a los destinatarios, casi ocho de cada diez contenidos (80%) se dirigieron a un grupo diana y no a personas concretas, lo que confirma el carácter estructural del fenómeno. Los colectivos más atacados fueron las personas originarias del norte de África (35%), las personas africanas y afrodescendientes (24%), las personas inmigrantes en sentido genérico (21%) y las personas musulmanas (en torno al 20%); también destacó, por su vulnerabilidad, el discurso dirigido a los niños, niñas y adolescentes no acompañados (5%). Por tipología, predominaron los mensajes que deshumanizan o degradan gravemente (37%), seguidos del descrédito basado en atributos personales (32%), la incitación a la violencia mediante amenazas directas o indirectas (29%), la presentación del grupo como amenaza (22%) y la incitación a la expulsión del colectivo (15%).

Resulta especialmente pertinente para el debate sobre la detección algorítmica la dimensión expresiva del discurso. Aunque el lenguaje agresivo explícito sigue siendo mayoritario (53%), el discurso discriminatorio no agresivo (36%) y, sobre todo, el tono irónico o sarcástico (11%, cuatro puntos más que el año anterior) ganan terreno; estas formas codificadas, disfrazadas de humor o crítica, son las más difíciles de identificar tanto socialmente como por los sistemas automatizados, lo que favorece su permanencia y normalización. Por último, buena parte de esta hostilidad se articula en torno a episodios prototípicos, encabezados por la vinculación entre inmigración e inseguridad

ciudadana (34%), de los cuales en torno al 40% se basa en percepciones, bulos o hechos descontextualizados, si bien un 21% del discurso no responde a ningún acontecimiento concreto, lo que confirma una base estructural de prejuicio que opera al margen de la coyuntura (OBERAXE, 2025).

En conjunto, estas cifras dibujan un fenómeno de gran amplitud que solo aflora de forma muy parcial en las estadísticas oficiales. Frente a los 2.870 contenidos de odio que el OBERAXE detectó y notificó a las plataformas en 2024 (OBERAXE, 2025), el Sistema Estadístico de Criminalidad registró únicamente 283 delitos e incidentes de odio cometidos a través de internet y redes sociales en 2025 (Ministerio del Interior, 2025). El contraste entre ambas magnitudes ilustra la dimensión de la cifra del odio digital: solo una fracción mínima del discurso hostil que circula en las redes, y que en su mayoría ni siquiera es retirado por las propias plataformas, llega a denunciarse y a cristalizar como infracción penal, pese a compartir con la criminalidad registrada sus mismos rasgos dominantes, la motivación racista y xenófoba y el protagonismo de las personas de origen norteafricano. Esta brecha confirma que el entorno digital no solo amplifica el discurso de odio, sino que, a través de unas dinámicas de moderación opacas y desiguales, contribuye a que buena parte de él permanezca invisible para la respuesta institucional.

4.2. Estudio de casos: el racismo digital en acción

Los estudios de casos constituyen, junto con el análisis estadístico, una herramienta metodológica de primer orden para el análisis del racismo digital, en la medida en que permiten descender desde el plano de las categorías abstractas hasta la concreción de los mecanismos reales mediante los cuales el odio racial se propaga, se amplifica y se normaliza en las redes sociales. Los casos ofrecen una comprensión de los procesos en juego atendiendo a quién habla, desde qué posición, con qué argumentos y con qué efectos sobre las personas y colectivos afectados. Los tres episodios que se examinan a continuación, el caso Vinícius Jr., el caso Badalona y el caso Mocejón, han sido seleccionados porque ilustran con particular claridad los patrones recurrentes del racismo digital en España en los que se evidencia la velocidad de propagación de los mensajes de odio, la participación de figuras públicas en su difusión, la instrumentalización de la desinformación para vincular criminalidad e inmigración y la insuficiencia de los mecanismos de moderación para contenerlos. No se trata, en ningún caso, de episodios excepcionales, más bien son ejemplos de una dinámica estructural que

se repite con regularidad (Solanes y Hernández, 2024).

El caso Vinícius Jr. ha puesto sobre la mesa la relevancia y gravedad de unos hechos que parecían pasar desapercibidos. Ha habido innumerables escenarios y partidos en los que el público ha proferido insultos racistas a jugadores de la primera y segunda división del fútbol español, pero la cobertura mediática de ninguno ha sido comparable a la que este episodio generó. Durante la temporada 2022-2023, el jugador había sufrido insultos racistas por parte de aficionados en el estadio Metropolitano de Madrid en septiembre de 2022, en Valladolid en diciembre de ese mismo año, en Mallorca en febrero de 2023 y en el estadio Camp Nou un mes después. Llegada la trigesimoquinta jornada del campeonato de liga el 21 de mayo de 2023, el Real Madrid se enfrenta al Valencia CF en el estadio de Mestalla, un encuentro en el que Vinícius Jr. es objeto de insultos racistas por parte de personas que él mismo identifica en la grada sur, si bien las agresiones verbales habían comenzado ya antes incluso de entrar en el propio estadio (Mercadal, 2024).

Según señala el acta del árbitro principal, De Burgos Bengoetxea, en el minuto 73 un espectador desde la grada sur se dirigió al jugador número 20 del Real Madrid CF realizando insultos racistas, por lo que se activó el protocolo de racismo, avisando al delegado de campo para que hiciesen el correspondiente aviso por megafonía. El encuentro estuvo detenido hasta que dicho anuncio se emitió por la megafonía del estadio (Comité Técnico de Árbitros, 2023). El racismo vivido en el estadio se trasladó a las redes sociales ese mismo día, tanto por el incidente cubierto en directo por la televisión y la prensa deportiva como por el mensaje que el jugador publicó en su perfil de X, en el que señalaba que el campeonato que alguna vez fue de Ronaldinho, Ronaldo, Cristiano y Messi ahora es el de los racistas, describiendo España como una nación hermosa que ha aceptado exportar al mundo la imagen de un país racista.

El I Informe sobre el estado de las migraciones y la convivencia intercultural en España de Convive-Fundación Cepaim ofrece un análisis de las respuestas en redes sociales al incidente. Entre sus resultados principales, destaca que el 70% de los mensajes analizados en X sobre el caso Vinícius Jr. en las veinticuatro horas posteriores a los hechos trataban el tema del racismo. El 59,32% de los perfiles analizados apoyaron al jugador y no se identificaron discursos de violencia extrema, lo cual puede indicar tanto su ausencia como que fueron denunciados y retirados tempranamente (Soriano et al., 2023, p. 17).

Entre los comentarios que contienen violencia ofensiva contra Vinícius Jr. (un

23,38% del total analizado) se hallan algunos abiertamente racistas y otros que justifican los insultos por el comportamiento previo inadecuado del jugador, culpabilizando así al futbolista del racismo sufrido. Otros usuarios expresaban que las quejas del jugador no eran relevantes y que estaba exagerando la situación. En ocasiones se llega al extremo de afirmar que la existencia de insultos racistas no implica que haya racismo. Villar, Sánchez y Pulido (2024, pp. 12-13) destacan que los periódicos deportivos dedican cada vez más espacio a abordar el racismo en el fútbol profesional, aunque todavía se le otorga una cobertura menor en comparación con la cantidad de contenido que publican diariamente, dando a entender que no parecen apreciar la existencia de un verdadero problema de racismo en el deporte español. En este caso concreto, el 54,3% de los contenidos que mencionan explícitamente el término "racismo" lo hacen asociado a este jugador y el 33,2% se centra específicamente en el incidente de Mestalla.

Por otra parte, el caso Badalona está relacionado con el principal episodio prototípico señalado por el OBERAXE como generador de discurso del odio: la inseguridad ciudadana. El episodio, relacionado con la violación grupal de una menor por parte de otros menores en un centro comercial de Badalona en noviembre de 2022 y conocido públicamente meses después, provocó una fuerte reacción pública marcada por el racismo y la xenofobia. Aunque la nacionalidad de los agresores no fue revelada inmediatamente, en las redes sociales rápidamente comenzaron a circular especulaciones sobre su origen extranjero. Diversas publicaciones sugirieron que el hecho de que no se hubiera revelado la identidad de los agresores se debía a un supuesto intento de las autoridades y de los medios de comunicación de proteger a las personas migrantes que habrían sido las autoras de este crimen, proyectando la narrativa de que esta omisión informativa solo se produce cuando los culpables pertenecen a una minoría.

El informe de Convive-Fundación Cepaim recopiló 1.171 mensajes publicados en X, 224 (19%) de los cuales formaban parte de un debate en torno al racismo, a pesar de que en un inicio no se había hecho pública la nacionalidad o etnia de los involucrados. De los 224 mensajes sobre racismo, la mayoría (179, un 79,91% del total) fueron clasificados como "violencia ofensiva". Solo un mensaje fue categorizado como "violencia extrema", por hacer apología del exterminio nazi. Por otro lado, 23 tuits fueron clasificados como *upstander*, es decir, aquellos que responden al discurso del odio racista con una postura ética de rechazo. Los colectivos más estigmatizados fueron los musulmanes, los gitanos, los menores no acompañados y, en general, las personas

migrantes. En muchos casos, se utilizaba un lenguaje que asociaba la inmigración con un incremento en la criminalidad, reforzando estereotipos negativos.

Se hicieron alusiones culturales para reforzar estas ideas; incluso se acuñó el término "elefantes" para referirse a un grupo identitario vinculado a los agresores, en alusión a un artículo de opinión que evitaba señalar su procedencia. Este término se convirtió en un código entre ciertos usuarios para aludir a la nacionalidad o etnia de los agresores sin mencionarla directamente. Además, algunas publicaciones incluían datos o supuestas estadísticas sobre el aumento de la delincuencia en relación con la inmigración, aunque no se proporcionaban fuentes fiables para validar estas afirmaciones. Este tipo de contenido, que utiliza la apariencia de datos empíricos para legitimar discursos racistas, es especialmente peligroso porque puede parecer más creíble para ciertos sectores de la población. Resulta significativo que solo el 10% de los mensajes analizados tuvieran una posición activa frente al racismo (Soriano Miras et al., 2023), lo que refleja la asimetría existente en las redes sociales entre la producción y el combate del discurso del odio.

Finalmente, el caso Mocejón guarda similitudes con el anterior en cuanto a que se trata de un episodio prototípico asociado a la inseguridad ciudadana, pero difiere en cuanto a la tipología delictiva al tratarse de un asesinato en el que, además, se atacó directamente desde las redes sociales a la familia de la víctima. La muerte del menor de once años en Mocejón, un municipio de la provincia de Toledo, el 18 de agosto de 2024, fue utilizada por determinados sectores para difundir bulos, desinformación y mensajes racistas y xenófobos en el ciberespacio, que continuaron a pesar de que la investigación policial descartó pronto la relación entre el crimen y la población migrante.

Desde el momento en que se conoció el suceso circularon teorías sobre la nacionalidad y la edad del autor, muchas de las cuales lo señalaban como magrebí o nacido en Rumanía, menor de edad y residente en un hotel que acoge a inmigrantes irregulares. El recientemente elegido eurodiputado Alvise Pérez publicó menos de cinco horas después de ocurridos los hechos un mensaje vinculando claramente el asesinato del menor con la llegada de personas africanas al pueblo semanas antes. Tres horas después publicó otro en el que afirmaba que el lunes 5 de agosto habían llegado al pueblo 50 africanos en un autobús, vinculando de forma directa la inseguridad ciudadana con la población migrante. También las cuentas anónimas tienen gran potencial para este tipo de difusión. Una de las primeras en hacerse eco de los hechos y cuyo mensaje fue visualizado más de un millón de veces antes de ser borrado afirmaba

que se había confirmado que el autor era un miembro de la comunidad norteafricana y pedía que España empezase a arder (Infantes, 2024).

Una investigación del discurso del odio racista en X arrojó 389 resultados coincidentes con una búsqueda de palabras clave durante el día 18 de agosto. Entre ellos, destacamos tres que, en conjunto, alcanzaron el millón y medio de visualizaciones en aquella red social. Todo esto sucedió en menos de veinticuatro horas, ya que el día 19 de agosto se procedió a la detención del presunto autor, de nacionalidad española y mayor de edad (Madrigal, 2024). Los ataques no se dirigieron solo contra el colectivo de personas migrantes, sino incluso hacia el portavoz de la propia familia de la víctima que pidió públicamente no criminalizar a nadie por su origen, a quien se le achacó tener "las manos manchadas por tener fotos en África", insinuando su posible responsabilidad por la muerte de su primo por publicar fotos de sus viajes como periodista por el continente africano.

En los casos de Badalona y Mocejón se observa un patrón recurrente: de forma inmediata, una vez conocido el suceso, se procede a especular con una autoría extranjera que se alimenta sobre la base de la falta de información al respecto en los medios de comunicación. El desconocimiento sobre la nacionalidad del autor se convierte automáticamente en certeza sobre su origen foráneo. Se pretende, de esta forma, asentar en el imaginario colectivo la idea de que la criminalidad está vinculada a la persona migrante. Contribuyen activamente a esta dinámica no solo personas anónimas, sino también representantes políticos y perfiles con muchos seguidores que centran su actividad en encender el fuego y propagar información falsa e interesada, siempre reforzada con elevadas dosis de racismo y xenofobia. Se trata de una verdadera acción coordinada que pone de manifiesto la necesidad de instrumentos regulatorios más eficaces.

4.3. El III Plan de Acción de lucha contra los delitos de odio

La aprobación del III Plan de Acción de Lucha contra los Delitos de Odio, con vigencia entre 2025 y 2028, representa un hito significativo en la evolución de la política española frente a este tipo de conductas y evidencia la voluntad institucional de dar continuidad y profundidad a una estrategia que lleva ya tres ediciones consolidando su enfoque integral. Elaborado desde la Secretaría de Estado de Seguridad, a través de la Oficina Nacional de Lucha contra los Delitos de Odio (ONDOD), con la participación de los cuerpos policiales estatales y autonómicos, la Fiscalía de Sala contra los Delitos

de Odio y Discriminación, así como organizaciones y asociaciones del tercer sector, este tercer instrumento estratégico se construye sobre los avances de sus predecesores y responde a los nuevos desafíos que plantea la evolución del fenómeno, con especial atención al discurso del odio en el entorno digital (Ministerio del Interior, 2025). Conviene subrayar que se trata del único país de la Unión Europea que cuenta con una oficina nacional dedicada exclusivamente a los delitos de odio y con tres planes de acción sucesivos en la materia, lo que ha valido a España el reconocimiento expreso de organismos como la Agencia de los Derechos Fundamentales de la Unión Europea, la Comisión Europea y la Oficina de Instituciones Democráticas y Derechos Humanos de la OSCE.

Para comprender la envergadura del III Plan, conviene situarlo en la trayectoria de los dos anteriores. El I Plan de Acción (2019-2021) centró sus esfuerzos en la formación específica de las Fuerzas y Cuerpos de Seguridad del Estado y en la sensibilización institucional, estableciendo los fundamentos de una respuesta coordinada frente a los delitos de odio en España. El II Plan (2022-2024) desplazó el foco hacia la atención a las víctimas y la colaboración interinstitucional, articulándose en ocho líneas estratégicas, quince objetivos y un total de ochenta y seis medidas, de las cuales se cumplieron ochenta y cinco. Entre sus logros más relevantes destacan la creación y consolidación de unidades especializadas en el seno de las Fuerzas y Cuerpos de Seguridad del Estado: los Equipos de Extremismo Violento y Odio (EVO) en la Policía Nacional y los Equipos de Respuesta a los Delitos de Odio (REDO) en la Guardia Civil, tanto a nivel central como periférico dentro de las estructuras de Información. Asimismo, la ONDOD participó como socio en proyectos financiados por la Comisión Europea orientados a la mejora de las capacidades de detección del discurso del odio en línea, como el proyecto REAL UP, el proyecto CISDO para la cooperación interpolicial y social contra los delitos de odio, y el proyecto SCORE en el ámbito del deporte y la lucha contra el racismo. También se avanzó en la atención a las víctimas con la publicación en julio de 2023 de la Guía de actuación de asistencia a víctimas de delitos de odio que presentan síntomas relacionados con trauma y factores de estrés, elaborada en colaboración con la Sociedad Española de Especialistas en el Trastorno de Estrés Postraumático, y se lideró y coordinó en diciembre de 2023, a propuesta de la Comisión Europea, el Día de Acción Conjunta de la UE sobre la Lucha Contra los Delitos de Odio, en colaboración con las Fuerzas y Cuerpos de Seguridad, el Centro de Inteligencia contra el Terrorismo y el Crimen Organizado y Europol.

El III Plan se articula en seis líneas de acción, quince objetivos y ciento nueve medidas, con una dotación económica total de 1.432.570 euros, cifra notablemente superior a la de los planes anteriores. La primera línea de acción se orienta a mejorar el apoyo a las víctimas y su conocimiento sobre los recursos disponibles en las Fuerzas y Cuerpos de Seguridad y en otras instituciones públicas y privadas; la segunda busca prevenir la comisión de los delitos y discursos de odio mediante el refuerzo de las relaciones entre los distintos actores implicados y la intensificación de los esfuerzos para minimizar la propagación del odio en internet; la tercera profundiza en la formación, sensibilización y concienciación de los miembros de las Fuerzas y Cuerpos de Seguridad, con especial atención a las nuevas promociones; la cuarta incentiva la participación y colaboración con el tercer sector en la prevención y lucha contra los delitos de odio; la quinta impulsa el desarrollo de un sistema de seguimiento y coordinación de las investigaciones y la generación de conocimiento útil para la prevención; y la sexta, a la que corresponde la parte más relevante del presupuesto, un total de 1.100.000 euros, persigue mejorar los recursos personales y materiales de los equipos EVO y REDO, creando estructuras centrales operativas especializadas en la persecución de los delitos de odio.

Las medidas del III Plan relativas al discurso del odio en línea revisten una importancia capital. En el ámbito de la investigación policial, el Plan contempla la constitución de un Grupo Operativo Central para la investigación de los delitos de odio cometidos en internet y redes sociales, con capacidad para actuar cuando el delito supera el ámbito territorial de los grupos EVO y REDO periféricos o cuando se desconoce el lugar de emisión del discurso. Se plantea asimismo el impulso al estudio y exploración de tecnologías emergentes en el ámbito de la inteligencia artificial para identificar y combatir el discurso de odio en línea desde un enfoque ético y responsable, así como la potenciación de los *Referral Action Days* para reportar contenidos de odio a los prestadores de servicios en línea, el impulso al trabajo conjunto entre la Oficina de Coordinación de Ciberseguridad y la ONDOD, y el establecimiento de reuniones periódicas con la Comisión Nacional de los Mercados y la Competencia como Punto Focal Nacional en el marco del Reglamento Europeo de Servicios Digitales. Es también relevante la propuesta de desarrollar herramientas automatizadas en colaboración con universidades para la detección e investigación de discursos de odio en redes sociales, así como el uso de análisis predictivo a partir de datos públicos para identificar patrones de comportamiento susceptibles de derivar en delitos o discursos de odio. Esta apuesta

por la colaboración entre las fuerzas de seguridad y la comunidad científica constituye un paso especialmente significativo, habida cuenta de la brecha que históricamente ha existido entre ambos ámbitos en el tratamiento de esta problemática.

El Plan también pone el acento en la formación especializada, con medidas que van desde la revisión y actualización periódica de los contenidos sobre delitos de odio en los planes de estudio de acceso y ascenso en las Fuerzas y Cuerpos de Seguridad, hasta el fomento de cursos universitarios para personal de los equipos EVO y REDO, pasando por la potenciación de la formación TAHCLE mediante un sistema de formación en cascada. Merece igualmente destacarse la propuesta de que los Equipos EVO de Policía Nacional y REDO de Guardia Civil asuman específicamente las investigaciones por delitos y discursos de odio cometidos en el ámbito deportivo, particularmente en el fútbol, estableciendo contactos directos con todos los operadores y entidades afectadas. Este extremo conecta directamente con el caso Vinícius Jr., analizado anteriormente, que reveló la necesidad de una respuesta policial y judicial rápida, coordinada y especializada en el ámbito del racismo en el deporte.

El III Plan reconoce explícitamente la gravedad de la infradenuncia como uno de los obstáculos estructurales más persistentes en la respuesta institucional a los delitos de odio (Pina et al, 2022). Muchas víctimas no son conscientes de haber sufrido un delito de estas características, otras desconfían del sistema o temen las consecuencias de denunciar, y sin denuncia no hay reparación ni condena posible. Para combatir este fenómeno, el Plan incluye campañas de sensibilización para que las víctimas conozcan sus derechos y los recursos disponibles tanto en espacios digitales como en puntos físicos, la incorporación en el atestado policial de información sobre recursos disponibles para la víctima, la mejora de la aplicación ALERTCOPS con una tarjeta específica sobre delitos de odio y el impulso a evaluaciones de riesgo de reiteración victimal. En este sentido, el Plan contempla además el desarrollo de apoyos de resiliencia comunitaria principalmente en áreas con alta presencia de incidentes o delitos de odio, donde se puedan ofrecer apoyo y asesoramiento psicológico, legal y social a las víctimas, así como programas de mentoría y apoyo entre pares.

Junto a estas medidas orientadas a las víctimas y a la investigación policial, el III Plan incorpora una dimensión preventiva de alcance más amplio. Se prevé el fortalecimiento de la educación y sensibilización pública en los distintos niveles educativos, incluyendo la organización de talleres de prevención y sensibilización en centros educativos y universidades, y el potenciamiento de las actuaciones formativas

sobre delitos y discursos de odio impartidas al alumnado. Asimismo, se impulsa la creación de la figura de los embajadores contra el odio, personas influyentes en sus comunidades como líderes religiosos, figuras deportivas o artistas, que realicen acciones de sensibilización y campañas contra los delitos de odio amplificando el mensaje hacia audiencias a las que los instrumentos institucionales clásicos no siempre llegan. Finalmente, la previsión de una Comisión de Seguimiento que se reunirá con carácter ordinario una vez al semestre para valorar el grado de cumplimiento del Plan, con la participación de representantes del tercer sector y de la Fiscalía especializada, constituye un mecanismo de rendición de cuentas que, si funciona como está previsto, puede contribuir a superar la brecha entre el diseño formal de las políticas públicas y su implementación efectiva sobre el terreno.

5. Conclusiones: propuestas de políticas públicas frente al racismo digital

Las claves conceptuales permiten entender que el racismo y la xenofobia en las redes sociales es un fenómeno particular que requiere una respuesta igualmente específica. A partir de ellas, puede plantearse un conjunto de propuestas concretas orientadas a combatir el discurso del odio racista en el entorno digital. Ambas dimensiones, la analítica y la propositiva, son inseparables. No es posible formular respuestas eficaces sin haber comprendido con precisión la naturaleza del problema al que se pretende hacer frente y, al mismo tiempo, un análisis que no desemboque en propuestas de acción corre el riesgo de quedarse en el plano puramente descriptivo. Se trata, en definitiva, de trasladar al terreno de las políticas públicas y de la acción institucional las conclusiones que el estudio del fenómeno hace posibles.

5.1. Claves para hacer frente al racismo y la xenofobia en las redes sociales

En primer lugar, conviene insistir en la idea de que, a pesar de que la raza es un constructo social que no se corresponde con una realidad biológica, el racismo sigue existiendo en formas diferentes a la tradicional asociada a la biología. En todas las maneras en las que el racismo se expresa subyace la idea de desigualdad de poder, que da paso a la discriminación y a la concentración excluyente de privilegios. El racismo actual es una construcción cultural como marcador de poder que adopta formas nuevas y más sutiles para legitimarse socialmente. No hay razas, por lo tanto, pero sí racismo, racistas y personas y grupos diana. Es racista quien, a pesar de la inexistencia de la raza como concepto aplicable a la especie humana, toma en consideración la categorización

biológica y racializada del ser humano, entendiendo que cada persona pertenece a una tipología concreta, siendo superior aquella de la que el racista forma parte.

En segundo lugar, el discurso del odio que se produce en las redes sociales es especialmente dañino. La lesión que produce el discurso del odio en las personas o grupos diana se recrudece si el medio empleado es una red social, puesto que las características inherentes de los contenidos que en ellas se publican ofrecen una vía idónea para su difusión inmediata a una audiencia potencialmente masiva. Además, el hecho de sentirse amparadas por el anonimato provoca una sensación de impunidad que permite a las personas racistas desinhibir su odio mediante expresiones públicas de hostilidad o incitación a la violencia hacia otros colectivos. El ánimo de lucro de las empresas que se hallan detrás de las plataformas de redes sociales está vinculado necesariamente a la maximización de la permanencia del usuario mediante la amplificación de contenidos polarizadores, incluidos los de naturaleza racista. Si el contacto amistoso y la comunicación son fundamentales para el entendimiento recíproco y este, a su vez, es esencial para eliminar la sensación de amenaza o peligro asociada a la alteridad, las redes sociales están propiciando justamente lo opuesto.

Un tercer elemento relevante es el hecho de que los episodios prototípicos son fuente recurrente de discurso de odio racista online. Dicho discurso se nutre principalmente de determinados eventos o situaciones que desatan la virulencia del odio y la hostilidad hacia determinados colectivos. El sesgo confirmatorio y las teorías de la conspiración incrementan dicho discurso, como los casos Badalona y Mocejón han evidenciado. Cuando los hechos no confirman los estereotipos racistas, la respuesta no es la rectificación sino la teoría conspirativa. Así, la ausencia de información sobre la identidad del presunto agresor se convierte en prueba de una supuesta protección institucional a la población migrante.

Una cuarta clave hace referencia al hecho de que las redes sociales han plataformizado la esfera pública y tienen una gran influencia en la conformación de la opinión pública. La plataformización de la esfera pública supone la existencia de un lugar virtual fácilmente accesible que, si bien desarrolla la función pública de contribuir a la formación de una opinión pública representativa de la voluntad popular, se halla sujeta a los intereses privados y económicos de las grandes compañías que las operan. En lugar de un espacio de interacción entre una ciudadanía que busca el bien común, nos encontramos con plataformas fragmentadas en nichos ideológicos impulsadas por intereses comerciales y diseñadas para capturar la atención mediante algoritmos que

priorizan el contenido que genera más reacciones, no necesariamente el más veraz o útil para el debate democrático. La polarización y fragmentación en los nichos ideológicos distorsionan la conformación de una voluntad popular deliberada, ya que, pone en riesgo el consenso real y libre entre sus participantes.

Como quinto elemento significativo es necesario insistir en que la libertad de expresión no es un derecho absoluto. La lucha contra el discurso de odio plantea un conflicto inevitable con este derecho fundamental. En Europa, y particularmente en España, se reconoce la necesidad de limitar este derecho cuando su abuso pone en peligro la dignidad de personas o colectivos en situaciones desfavorecidas o de vulnerabilidad. Los discursos del odio no solo impactan a las víctimas directas, sino que generan un ambiente de intimidación que desincentiva la participación de otras personas, reduciendo la calidad del diálogo público y la diversidad de opiniones. El principio de libertad de expresión que se busca proteger acaba, paradójicamente, minando las bases mismas de la deliberación democrática inclusiva. El antídoto frente al discurso del odio no es, sin embargo, exclusivamente la represión penal. Junto a ella, como última ratio, son imprescindibles también instrumentos de naturaleza educativa, institucional y regulatoria de las propias plataformas.

5.2. Propuestas para combatir el discurso del odio racista en las redes sociales

Las reflexiones precedentes nos colocan en situación de defender la limitación de la libertad de expresión para salvaguardar la dignidad de las personas y grupos diana del discurso del odio, así como la cohesión social y el pluralismo esencial en una democracia. Ahora bien, no todo discurso odioso es sancionable. Como advertíamos, no toda expresión de esa índole es penalmente relevante; solo cuando el mensaje desagradable alcanza un umbral lesivo determinado puede y debe entrar el derecho penal a reprimir comportamientos que dañan elementos esenciales reconocidos constitucionalmente como el pluralismo político, la libertad, la justicia y la igualdad. La dificultad para determinar qué contenido puede y debe ser sancionable pretende ser solventada a través del umbral propuesto por el Plan de Acción de Rabat, empleado por la Estrategia y Plan de Acción de Naciones Unidas.

El umbral del Plan de Rabat, empleado como mecanismo para detectar el momento en que una conducta expresiva de odio se convierte en ilegítima, opera sobre la base de seis criterios: el contexto en el que se emite el discurso, la identidad de la persona que lo realiza, la intención del mensaje, el contenido y la forma del discurso, la

magnitud de su difusión y la probabilidad e inminencia de daño. Sin embargo, cada uno de ellos presenta retos en su aplicación en el contexto del ciberespacio. El análisis del contexto exige un conocimiento profundo de las dinámicas sociopolíticas de cada lugar, ya que el impacto del mensaje puede variar considerablemente según el espacio geográfico. El elemento de difusión está más que superado en el caso de las redes sociales por el potencial que todo mensaje publicado en ellas tiene para llegar a cualquier usuario de la plataforma. El umbral del Plan de Rabat ofrece un marco de partida, pero resulta insuficiente en el contexto digital, donde el impacto puede ser significativo a pesar de no cumplir con todos sus criterios.

En el ámbito de la regulación de las plataformas digitales, la iniciativa Social Media 4 Peace de Naciones Unidas publicó en 2023 una serie de recomendaciones orientadas a las compañías tecnológicas entre las que se pueden destacar las siguientes: adaptar las políticas de sus plataformas a los estándares internacionales; diseñar algoritmos que tengan en cuenta la paz, la dignidad humana y el imperio de la ley como valores fundamentales y que restrinjan aquellos motivadores de la atención de las personas usuarias que sean contrarios a los objetivos de cumplimiento de derechos humanos; informar sobre el número de acciones relacionadas con cuentas y contenido relativo al discurso del odio; y establecer puntos de contacto locales para ofrecer apoyo a las víctimas (UNESCO, 2023). Los Principios Rectores sobre las empresas y los derechos humanos de Naciones Unidas estipulan el deber de las empresas de prevenir, mitigar y remediar las vulneraciones de derechos humanos que puedan provocar o a las que puedan contribuir con sus servicios. Asimismo, los principios recogidos en el informe sobre transparencia y responsabilidad en la era digital de la propia UNESCO (2021) abogan por que las compañías reconozcan expresamente su obligación de proteger los derechos humanos, sean transparentes en la elaboración de los términos y estándares que aplican y establezcan claramente los límites de lo que consideran un comportamiento aceptable.

En el ámbito español, el Protocolo para combatir el discurso de odio ilegal en línea coordinado por el OBERAXE y publicado en 2021 constituye un instrumento valioso. Este protocolo establece elementos para la identificación, notificación y bloqueo de contenidos de odio ilegal en internet, respetando al mismo tiempo los derechos fundamentales. Define a las "autoridades competentes" y el "punto de contacto" para el proceso de notificación y bloqueo de contenidos de odio, designando a la Unidad de Criminalidad Informática de la Fiscalía General del Estado como ese punto

de contacto. El protocolo establece asimismo mecanismos restaurativos alternativos en situaciones de discursos del odio que no ameriten una protección desde el ámbito penal. Para mejorar su eficacia, sería beneficioso que el papel del comunicante fiable sea optimizado al facilitar mayores recursos de capacitación y apoyo legal, que se clarifiquen los procedimientos específicos de los mecanismos restaurativos y que se cree una red de apoyo con recursos educativos para el público, fomentando una cultura de resolución pacífica y prevención del odio.

Junto a las medidas de carácter regulatorio, resulta imprescindible el fomento de la educación y la alfabetización digital. La ciudadanía no está educada en el entorno virtual, desconoce su funcionamiento más allá de la interfaz gráfica de usuario y tampoco ha tenido la oportunidad de formarse éticamente para actuar en espacios donde prima el anonimato y la sensación de impunidad. Una educación cívica que incorpore la dimensión digital resulta esencial para dotar a la ciudadanía de las herramientas necesarias para participar en la esfera pública digital de manera crítica, respetuosa y comprometida con los derechos humanos.

Además de la necesaria participación y colaboración de la sociedad civil, las organizaciones no gubernamentales y personal experto en derechos humanos para ofrecer respuestas efectivas desde el terreno, es esencial la actuación de los Estados y de las compañías prestadoras de servicios en internet y dueñas de las plataformas de las redes sociales, que tienen una responsabilidad compartida para luchar contra el discurso del odio racista en línea.

Del mismo modo, la diversidad en los equipos de diseño tecnológico es una condición necesaria para mitigar los sesgos algorítmicos que reproducen y amplifican las desigualdades existentes. Los algoritmos que reproducen estereotipos racistas no son fruto de la casualidad sino de unas condiciones de diseño en las que la diversidad de perspectivas es estructuralmente deficitaria. Equipos más diversos, auditorías externas obligatorias y estándares de no discriminación en los sistemas de inteligencia artificial son instrumentos necesarios, aunque no suficientes por sí solos.

El racismo digital no es una anomalía del sistema tecnológico que pueda resolverse con mejores filtros o más moderadores, es más bien una manifestación de desigualdades sociales preexistentes que encuentran en el entorno digital condiciones excepcionalmente favorables para su reproducción y amplificación. La respuesta, en consecuencia, no puede limitarse al plano técnico ni al jurídico, aunque ambos sean imprescindibles. Exige una intervención pública que interpele a las estructuras

económicas, políticas y culturales que sostienen el racismo en sus múltiples dimensiones, y que tenga la valentía de supervisar, con claridad y rigor, la existencia de actores con intereses, con poder y con responsabilidades que hasta ahora han ejercido de manera muy deficiente. La democracia digital que la ciudadanía merece aún está por construir.

7. Referencias.

Aguilar Idáñez, María José y Buraschi, Daniel (2016). Del racismo y la construcción de fronteras morales a la resistencia y el cambio social: la sociedad civil frente a las migraciones forzadas. *Servicios Sociales y Política Social*, 111, 29-44.

Aguilar Idáñez, María José y Buraschi, Daniel (2019). Racismo «democrático» y fronteras morales: ¿Cómo construir una ciudadanía insurgente? En Ángeles Solanes (dir.), *Discriminación, racismo y relaciones interculturales* (pp. 155-188). Thomson Reuters Aranzadi.

Alonso Sanz, Lucía (2017). La protección de los inmigrantes frente al discurso del odio. En Lucía Alonso y Víctor Jesús Vázquez (dirs.), *Sobre la libertad de expresión y el discurso del odio. Textos críticos* (pp. 241-262). Athenaica.

Anderson, Benedict (2006). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Aridor, Guy; Jimenez-Duran, Rafael; Levy, Ron y Song, Lulu (2023). The Economics of Social Media. *Journal of Economic Literature*, 62(4), 1422-1474. <https://doi.org/10.1257/jel.20221643>

Azani, Eitan; Koblenz-Stenzler, Lorena; Atiyas-Lvovsky, Liran; Ganor, Daphna y Meshlam, David (2020). The Development and Characterization of Far-Right Ideologies. En Eitan Azani et al. (eds.), *The Far Right: Ideology, Modus Operandi and Development Trends* (pp. 13-36). International Institute for Counter-Terrorism.

Baker, C. Edwin (2009). Autonomy and Hate Speech. En Ivan Hare y James Weinstein (eds.), *Extreme Speech and Democracy* (pp. 139-157). Oxford University Press.

Balkin, Jack M. (2018). Free Speech is a Triangle. *Columbia Law Review*, 118(7), 2011-2056.

Bandura, Albert (2002). Selective Moral Disengagement in the Exercise of

Moral Agency. *Journal of Moral Education*, 31(2), 101-119.
<http://dx.doi.org/10.1080/0305724022014322>

Barberá, Pablo (2020). Social Media, Echo Chambers, and Political Polarization. En Nathaniel Persily y Joshua A. Tucker (eds.), *Social Media and Democracy. The State of the Field and Prospects for Reform* (pp. 34-55). Cambridge University Press.

Belli, Luca y Zingales, Nicolo (2020). Platform value(s): A multidimensional framework for online responsibility. *Computer Law and Security Review*, 36, 105364.
<https://doi.org/10.1016/j.clsr.2019.105364>

Bizumic, Boris y Duckitt, John (2012). What Is and Is Not Ethnocentrism? A Conceptual Analysis and Political Implications. *Political Psychology*, 33(6), 887-909.

Bliuc, Ana-Maria; Faulkner, Nicholas; Jakubowicz, Andrew y McGarty, Craig (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, 87, 75-87.
<https://doi.org/10.1016/j.chb.2018.05.026>

Bratt, Christopher (2022). Is It Racism? The Belief in Cultural Superiority across Europe. *European Societies*, 24(2), 207-228.

Chen, Long; Chen, Jianguo y Chunhe, Xia (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64, 103060.

Clucas, Thomas (2020). «Don't Feed the Trolls»: Social Media and the Limits of Free Speech. En Sandra Polak y Daniel Trottier (eds.), *Violence and Trolling on Social Media*. Amsterdam University Press.

Comité Técnico de Árbitros (2023). Acta del partido celebrado el 21 de mayo de 2023, en Valencia, temporada 2022/2023. Disponible en:
https://historicoactas.rfef.es/actas/RFEF_CmpActa2?cod_primaria=1000144&CodActa=81607

DeNardis, Laura (2014). *The Global War for Internet Governance*. Yale University Press.

Farrell, Amy y Lockwood, Stacy (2023). Addressing Hate Crime in the 21st Century: Trends, Threats, and Opportunities for Intervention. *Annual Review of Criminology*, 6, 107-130. <https://doi.org/10.1146/annurev-criminol-030920-091908>

Fernández, Rodrigo; Adriaans, Ilke; Klinge, Tobias J. y Hendrikse, Reijer (2020). *Engineering Digital Monopolies. The Financialisation of Big Tech*. Centre for Research on Multinational Corporations.

- Floridi, Luciano (2024). *Ética de la inteligencia artificial*. Herder.
- FRA – European Agency for Fundamental Rights (2021). *Encouraging Hate Crime Reporting. The Role of Law Enforcement and Other Authorities*. Publications Office of the European Union.
- (2024). *Fundamental Rights Report 2024*. Publications Office of the European Union.
- Fuchs, Christian (2019). Karl Marx in the Age of Big Data Capitalism. En David Chandler y Christian Fuchs (eds.), *Digital Objects, Digital Subjects* (pp. 53-71). University of Westminster Press.
- Ganesh, Bharath (2018). The Ungovernability of Digital Hate Culture. *Journal of International Affairs*, 71(2), 30-49.
- Gascón Cuenca, Andrés (2016). *El discurso del odio en el ordenamiento jurídico español: su adecuación a los estándares internacionales de protección*. Cizur Menor: Thomson Reuters Aranzadi.
- Gongane, Vaishali U., Munot, Mukesh V. y Anuse, Alwin D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12 (1). DOI:10.1007/s13278-022-00951-3
- Habermas, Jürgen (1989). The Public Sphere: An Encyclopedia Article. En Stephen Eric Bronner y Douglas M. Kellner (eds.), *Critical Theory and Society. A Reader* (pp. 136-142). Routledge.
- Habermas, Jürgen (2023). *A New Structural Transformation of the Public Sphere and Deliberative Politics*. Polity Press.
- Han, Byung-Chul (2023). *La crisis de la narración*. Herder.
- Harari, Yuval Noah (2019). *21 lecciones para el siglo XXI*. Debate.
- Infantes Capdevilla, Guillermo (2024). La campaña de odio que relaciona el crimen de Mocejón con migrantes: actores, antecedentes y consecuencias. *Newtral* (22 de agosto de 2024). Disponible en: <https://www.newtral.es/campana-odio-crimen-mocejon-inmigrantes-menas-magrebies-bulo-desinformacion/20240822/>
- Jørgensen, Rikke Frank (2020). Rights Talk: In the Kingdom of Online. En Rikke Frank Jørgensen (ed.), *Human Rights in the Age of Platforms* (pp. 163-187). MIT Press.
- Karp, Alexander C. y Zamiska, Nicholas W. (2025). *La república tecnológica*.

Poder duro, pensamiento débil y el futuro de Occidente. Tenos.

Kemp, Simon (2024). *Digital 2024: Global Overview Report*. We Are Social y Meltwater.

Kim, Byungjoo (2020). Effects of Social Grooming on Incivility in COVID-19. *Cyberpsychology, Behavior and Social Networking*, 23(8), 519-525.

Klein, Adam (2017). *Fanaticism, Racism, and Rage Online: Corrupting the Digital Sphere*. Palgrave Macmillan.

Laub, Zachary (2019). *Hate Speech on Social Media: Global Comparisons*. Council on Foreign Relations.

Lea, Richard (2020). Google Photos still can't tell gorillas from people. *The Guardian*.

Madrigal, Agustín (2024). Investigación sobre discurso del odio racista en X durante el caso Mocejón. Disponible en: https://x.com/Madrigal_C/status/1825188484963029475

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina and Galstyan, Aram (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1-35. <https://doi.org/10.1145/3457607>

Menczer, Filippo and Hills, Thomas (2020). Understanding how algorithms and manipulators exploit our cognitive vulnerabilities empowers us to fight back. *Scientific American*, Diciembre, 54-60.

Mercadal, Javier (2024). «Qué pasó entre el Valencia y Vinicius: cronología de un conflicto», *Relevo* (28 de febrero de 2024). Disponible en: <https://www.relevo.com/futbol/liga-primera/paso-valencia-vinicius-cronologia-conflicto-20240228143240-nt.html>

Metzler, Hannah and Garcia, David (2023). Social Drivers and Algorithmic Mechanisms on Digital Media. *Perspectives on Psychological Science*, 19(5), 735-748. <https://doi.org/10.1177/17456916231185057>

Ministerio de Inclusión, Seguridad Social y Migraciones (2021). *Protocolo para combatir el discurso de odio ilegal en línea*. Secretaría de Estado de Migraciones.

Ministerio de Interior (2025). *III Plan de Acción de Lucha contra los Delitos de Odio 2025-2028*. Secretaría de Estado de Seguridad.

Noble, Safiya Umoja (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

OBERAXE (2025). *Informe anual de monitorización del discurso de odio en*

redes sociales 2024. Ministerio de Inclusión, Seguridad Social y Migraciones.

Oficina Nacional de Lucha contra los Delitos de Odio (2026). *Informe sobre la evolución de los delitos de odio en España 2025*. Ministerio del Interior.

Pasquale, Frank (2016). Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power. *Theoretical Inquiries in Law*, 17(2), 487-513.

Perkowitz, Sidney (2021). The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Schwarzman College of Computing*. DOI 10.21428/2c646de5.62272586

Pina Castillo, Víctor, Hernández Prados, María Ángeles y Galián Nicolás, Beatriz (2022). Los delitos de odio en España: análisis de la infradenuncia y propuestas de mejora. *Revista de Victimología*, 14, 67-92.

Russell, Alex (2023). YouTube Video Recommendations Lead to More Extremist Content for Right-Leaning Users, *UC Davis*.

Saito, Natsu Taylor (2021). Why Xenophobia? *Berkeley La Raza Law Journal*, 31, 1-25.

Sax, Ben; Greenblatt, Jonathan; Sheetz, Mike; Neufeld, Adam; Eisenstat, Yaël and Krapf, Lauren (2020). From Bad to Worse: Algorithmic Amplification of Antisemitism and Extremism. *ADL and Center for Technology and Society*.

Seeliger, Martin y Seignani, Sebastian (2022). A New Structural Transformation of the Public Sphere? An Introduction. *Theory, Culture & Society*, 39(4), 3-16.

Siegel, Lee (2008). *Against the Machine: How the Web is Reshaping Culture and Commerce and Why It Matters*. Spiegel & Grau.

Silva, Shelly y Kenney, Martin (2018). Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon*, 55(1-2), 9-37.

Solanes Corella, Ángeles y Hernández Moreno, Nacho (2024). *Formas de combatir el racismo en las redes sociales*. Tirant lo Blanch.

Solanes Corella, Ángeles y Hernández Moreno, Nacho (2025). Racism in the Digital Age: The Impact of Social Media Algorithms on Public Discourse. *The Age of Human Rights Journal*, 25, e9603. <https://doi.org/10.17561/tahrj.v25.9603>

Soriano Miras, Rosa María; De Botton Fernández, Lena; Burgés Freitas, Ana y Granda Ortells, Lucía (2023). *I Informe sobre el estado de las migraciones y la convivencia intercultural en España. Estudio de casos: racismo y xenofobia en las redes sociales digitales en España*. Convive-Fundación Cepaim.

Susskind, Richard (2025). *Descifrando la IA. Cómo usarla sin perder el juicio*. Dykinson.

Tapia Ballesteros, Patricia (2021). El discurso de odio del art. 510.1.a) del Código Penal español: la ideología como un Caballo de Troya entre las circunstancias sospechosas de discriminación, *Política Criminal: Revista Electrónica Semestral de Políticas Públicas en Materias Penales*, 16, (31), 284-320.

UNESCO (2021). *Letting the Sun Shine In: Transparency and Accountability in the Digital Age*. UNESCO.

UNESCO (2023). *Social Media 4 Peace. Local Lessons for Global Practices*. UNESCO.

Vaidhyanathan, Siva (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford University Press.

Villar-Cirujano, Ernesto; Sánchez de la Nieta, Miguel Ángel y Pulido Zaragoza, Julio (2024). El tratamiento del racismo en la prensa deportiva online española: el caso Vinicius. *ICONO 14. Revista Científica de Comunicación y Tecnologías Emergentes*, 22(1). <https://doi.org/10.7195/ri14.v22i1.2077>

Waldron, Jeremy (2012). *The Harm in Hate Speech*. Harvard University Press.

Weaver, Simon (2011). *The Rhetoric of Racist Humour: US, UK and Global Race Joking*. Ashgate.

Zuboff, Shoshana (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.