

Modelos de moderación de contenido, mecanismos de autorregulación y dinámicas de autolegitimación

Content moderation models, self-regulation mechanisms, and self-legitimation dynamics

Rodrigo Ardisson de Souza¹
ORCID: 0009-0002-0640-5052
CIDE, México

Dardo Emanuel Neubauer²
ORCID: 0009-0001-9701-0509
PUEDJS-UNAM, México

[Recibido: 21/03/2025 - Aceptado para su publicación: 19/05/2026]
<http://dx.doi.org/10.32870/Pk.a16n31.933>

Resumen

La investigación aborda los cambios recientes en la moderación de contenidos de Meta y su posible repercusión en la libertad de expresión y la protección de derechos fundamentales en el entorno digital. Se examinan principalmente el fin de la colaboración con verificadores externos y la relajación de las normas comunitarias en áreas sensibles como género, migración y discriminación, a la luz de diversos marcos reglamentarios en el mundo y aquellos preceptos teóricos que ponen en discusión la responsabilidad de las plataformas digitales en materia de moderación de contenido. A partir de este análisis, se constata una tendencia hacia la desregulación y la transición hacia la verificación

por parte de los usuarios para denunciar o verificar el contenido, lo que podría facilitar la proliferación de discursos de odio y la difusión de campañas de desinformación. El alineamiento de Meta con la nueva administración de Donald Trump agrava la tensión entre las *big tech* y los Estados, evidenciando el desequilibrio entre la autonomía corporativa y las prerrogativas estatales para salvaguardar los derechos de la ciudadanía. Los hallazgos reafirman la necesidad de contar con mecanismos de auditoría independientes y regulaciones coherentes, a fin de conciliar la innovación tecnológica con la protección efectiva de la esfera pública y los valores democráticos.

Palabras clave: Libertad de expresión; medios sociales; gobernanza de Internet; plataforma digital; discurso de odio.

Abstract

This research examines recent changes in Meta's content moderation policies and their potential impact on freedom of expression and the protection of fundamental rights in the digital environment. The study focuses primarily on the termination of collaborations with external fact-checkers and the relaxation of community guidelines in sensitive areas such as gender, migration, and discrimination. These developments are analyzed in light of various regulatory frameworks worldwide, as well as theoretical perspectives that challenge the responsibility of digital platforms in content moderation. The analysis reveals a trend toward deregulation and a shift to user-driven content

verification and reporting, which may facilitate the spread of hate speech and disinformation campaigns. Furthermore, the alignment of Meta with the new administration of Donald Trump exacerbates tensions between big tech companies and states, highlighting the imbalance between corporate autonomy and state prerogatives in safeguarding citizens' rights. Within this context, the findings underscore the urgent need for independent auditing mechanisms and coherent regulations to reconcile technological innovation with the effective protection of the public sphere and democratic values.

Keywords: *Freedom of expression; social media; Internet governance; digital platform; hate speech.*

Introducción

El 7 de enero de 2025, Mark Zuckerberg, director ejecutivo de Meta anunció la eliminación de los verificadores externos y la modificación de sus normas comunitarias, en lo que representó un giro sustancial en sus políticas de moderación de contenido (Kaplan, 2025). Estas modificaciones transforman la forma en que los usuarios interactúan con estas plataformas y podrían tener implicaciones profundas en la configuración de las dinámicas de desinformación, discursos de odio y violaciones a los derechos humanos. Hasta ese momento, Meta colaboraba con entidades externas especializadas en la verificación de hechos (*fact-checkers*) y contaba con sistemas automatizados internos diseñados para bloquear o limitar la difusión de contenidos violentos y discriminatorios. No obstante, con su anuncio, la compañía decidió poner fin a estas alianzas y adoptar un nuevo sistema de “notas de la comunidad”, inspirado en el modelo implementado por Elon Musk en la plataforma X.

Además, Meta redujo drásticamente su intervención en la moderación de contenidos, limitándola únicamente a casos considerados de “alta gravedad”, como el terrorismo, la explotación sexual infantil y el tráfico de drogas. El resto quedará, en gran medida, en manos de los usuarios, quienes deberán reportar los contenidos inapropiados a través de denuncias. Al reducir su papel activo en la curaduría del contenido, Meta transfiere la responsabilidad a la comunidad de usuarios y, al mismo tiempo, redefine la estructura de poder dentro de su ecosistema digital. La empresa se desliga de la carga operativa y de costos asociados a la moderación, pero sin establecer mecanismos claros de responsabilidad sobre los efectos que esta

desregulación puede generar. De este modo, la supuesta descentralización de la moderación no implica necesariamente una democratización del espacio digital; puede derivar, en cambio, en una fragmentación de la gobernanza de contenidos, donde la falta de intervención directa por parte de la plataforma podría fortalecer dinámicas de impunidad frente a la proliferación de discursos nocivos.

Zuckerberg justificó la medida aludiendo a la necesidad de “restaurar” o “ampliar la libertad de expresión” y alegando que los filtros de control “cometen errores y censuran contenido legítimo”. No obstante, diversos analistas (Direitos na Rede, 2025; IRISBH, 2025) advierten que esta supuesta defensa de la libre expresión resta responsabilidades básicas de la empresa y podría, en realidad, favorecer la proliferación de discursos discriminatorios y violentos. El hecho de prescindir de la verificación independiente y relajar los criterios de moderación y las normas comunitarias de la compañía, amenaza con socavar la calidad de la información que circula en línea, en especial en un contexto geopolítico cada vez más convulso y polarizado. La experiencia de crisis globales recientes, como la pandemia, demostró que la ausencia de mecanismos efectivos de control promueve la difusión masiva de desinformación con repercusiones potencialmente graves para la salud pública y el tejido democrático.

El alineamiento entre las big techs y el gobierno de Donald Trump en su segundo mandato marca un punto central en la relación entre el poder político y las grandes corporaciones tecnológicas. El fenómeno aquí analizado no es solo una redefinición en la relación entre corporaciones tecnológicas y Estado, sino la consolidación de un modelo híbrido donde las big techs dejan de actuar únicamente como estados paralelos con normas propias y sistemas algorítmicos de control para integrarse formalmente al poder político. Esta integración profundiza la captura regulatoria, transformando desde adentro la estructura estatal en función de los intereses corporativos. Así, figuras como Musk, Zuckerberg o Bezos, tradicionalmente asociadas al discurso de neutralidad y libre mercado, revelan ahora una nueva dinámica donde la soberanía y la gobernanza quedan redefinidas por la alianza explícita entre la industria digital y el poder político tradicional.

Metodología

La presente investigación parte de un enfoque jurídico y analítico basado en el análisis crítico-comparativo de marcos normativos internacionales, jurisprudencia relevante y doctrina especializada en materia de moderación de contenidos digitales, autorregulación corporativa y protección de derechos fundamentales. La metodología adoptada se fundamenta en la revisión y sistematización de documentos normativos, pronunciamientos judiciales, comunicados oficiales y literatura académica que

permitan una interpretación rigurosa y objetiva de los recientes cambios en las políticas de moderación de la plataforma Meta.

Inicialmente, se realizó una revisión normativa comparada que comprende los marcos regulatorios clave de Estados Unidos (Sección 230 del Communications Decency Act), la Unión Europea (Digital Markets Act y Digital Services Act), América Latina (Marco Civil de Internet de Brasil y el T-MEC entre México, EE.UU. y Canadá), con especial énfasis en identificar los puntos de convergencia y divergencia que permitan comprender las implicaciones jurídicas derivadas de los cambios anunciados por Meta. Esta comparación permitió visibilizar los diferentes niveles de responsabilidad exigidos a las plataformas digitales en contextos jurídicos distintos, aportando claridad sobre cómo los recientes cambios de Meta interactúan con estas obligaciones.

Además, se incorporó un análisis centrado en resoluciones y opiniones vinculantes o influyentes emitidas por diferentes organismos destacando decisiones relacionadas con derechos fundamentales, libertad de expresión y responsabilidad corporativa en el entorno digital. En particular, se estudió el papel del Oversight Board de Meta, atendiendo a sus límites operativos y a la repercusión real de sus decisiones, así como la interpretación doctrinal sobre la efectividad y legitimidad de estos mecanismos autorreguladores.

Para complementar el análisis se incorporaron fuentes documentales provenientes de organismos internacionales y entidades especializadas en verificación de contenidos y sociedad civil, como la International Fact-Checking Network (IFCN), UNESCO, ARTICLE 19 y reportes de organizaciones no gubernamentales especializadas en derechos digitales, con el fin de evaluar críticamente el impacto potencial de la transición hacia modelos menos intervencionistas en moderación de contenidos.

Finalmente, la metodología también consideró la revisión sistemática de literatura académica reciente sobre gobernanza digital (Gorwa, 2019, 2024; Flew, 2021), autorregulación (Klonick, 2018; Suzor, 2019; Douek, 2022), soberanía digital (Couture y Toupin, 2019; Pohle y Thiel, 2020; Jiang y Belli, 2025) y teoría del discurso (Siapera y Viejo-Otero, 2021; Habermas, 2025; Magalhães et al., 2026), permitiendo construir una perspectiva teórica sólida que diera sustento al análisis crítico de las decisiones adoptadas por Meta. Este enfoque permitió argumentar de manera objetiva y fundamentada sobre las consecuencias jurídicas y sociales de la desregulación en la moderación de contenidos y sus posibles efectos en términos de protección de derechos humanos y preservación del espacio público digital.

En conjunto, esta metodología proporciona un marco analítico robusto y jurídicamente sólido que permite evaluar objetivamente las recientes modificaciones

implementadas por Meta, contribuyendo a la identificación de tendencias generales y desafíos regulatorios emergentes en materia de gobernanza digital.

Desarrollo

El doblete: Estado oficial y Estado paralelo

Las plataformas digitales funcionan como infraestructuras para la deliberación pública, la economía y la seguridad; se consolidan, además, como estructuras de gobierno con capacidad de incidir directamente en la formulación de políticas. La creciente permisividad hacia discursos extremistas y la eliminación progresiva de mecanismos de moderación de contenido reflejan un giro estratégico en la gestión de las plataformas y una alineación ideológica con los marcos discursivos que sustentan el nuevo gobierno. La lógica de la “libertad de expresión” promovida por Musk y Trump se instrumentaliza como un mecanismo de desregulación que desdibuja las fronteras entre lo público y lo privado, permitiendo que estas corporaciones operen con autonomía y adquieran una función estructural en la definición del orden político contemporáneo.

Este modelo de gobernanza digital compartida o coregulada cuestiona la capacidad de los Estados de ejercer controles efectivos sobre las plataformas y sobre los riesgos asociados a la concentración de poder en manos de actores privados con intereses propios. La convergencia entre el poder político y las big techs, lejos de representar una mera reconfiguración administrativa, señala la consolidación de un modelo de soberanía digital donde las grandes corporaciones tecnológicas median la esfera pública e intervienen activamente en la producción del marco normativo que rige la sociedad digital.

El progresivo abandono de los sistemas tradicionales de moderación, que incluían verificadores independientes y controles automatizados estrictos, donde mínimamente las plataformas entienden su papel de empresas privadas, y tratan de mitigar riesgos, aunque de manera bastante insuficiente, ha dado lugar a modelos más descentralizados y ambiguos. Herramientas como las Community Notes en X (Twitter), diseñadas bajo la premisa de permitir la corrección colaborativa de información, trasladan la responsabilidad del control de contenidos hacia los propios usuarios, generando nuevas asimetrías en la regulación del discurso público. A su vez, estructuras como el Oversight Board de Meta buscan proyectar una imagen de independencia y accountability, pero en la práctica refuerzan la capacidad de la empresa para legitimar sus propias decisiones bajo un manto de autorregulación institucionalizada. Este desplazamiento hacia modelos de moderación menos intervencionistas y más descentralizados implica una reconfiguración del papel de las

plataformas en la gobernanza digital y transforma la forma en que se articula la soberanía informacional.

La creciente externalización de la moderación de contenidos a mecanismos comunitarios o a órganos corporativos con autonomía limitada permite que las big techs mantengan el control sobre las reglas del espacio digital sin asumir plenamente la responsabilidad por los efectos de la desinformación y la amplificación de discursos extremos. Esa convergencia entre el poder político y las big techs no es solo una cuestión de captura regulatoria o influencia en la formulación de políticas estatales. Se trata de una transformación estructural en la que las infraestructuras digitales dejan de ser meros espacios de mediación y se convierten en entornos normativos en sí mismos, estableciendo criterios de visibilidad, acceso y validación de la información.

La estrategia de Meta de reducir su rol en la moderación de contenidos ocurre en un contexto regulatorio fragmentado. En EE.UU., la Sección 230 del Communications Decency Act mantiene la inmunidad para plataformas digitales respecto a contenidos generados por usuarios, aunque esta doctrina enfrenta críticas crecientes por facilitar la desinformación y el discurso de odio. En Europa, el Digital Markets Act (DMA) y el Digital Services Act (DSA) establecen un régimen más exigente; el DMA impone obligaciones específicas a los "gatekeepers" sobre transparencia algorítmica y competencia digital, el DSA fortalece las responsabilidades sobre la moderación de contenidos ilícitos, exigiendo mecanismos claros de respuesta y transparencia en las decisiones editoriales. En América Latina, el Marco Civil de Internet brasileño adopta una responsabilidad condicionada, que garantiza neutralidad y libertad de expresión, pero impone obligaciones puntuales ante contenido ilegal, particularmente bajo órdenes judiciales. Finalmente, el T-MEC entre México, EE.UU. y Canadá se alinea parcialmente con la doctrina estadounidense, limitando la responsabilidad directa de intermediarios digitales, aunque permite regulaciones nacionales más estrictas, reflejadas en recientes iniciativas en Canadá y México. Estas normativas señalan un desplazamiento progresivo hacia mayores niveles de rendición de cuentas y diligencia debida para las plataformas digitales. La respuesta institucional a los cambios anunciados por Meta en enero de 2025 confirmó esa tensión en términos concretos. La Comisión Europea rechazó de inmediato las acusaciones de "censura" formuladas por Zuckerberg: el portavoz Thomas Regnier aclaró que si Meta desea adoptar el modelo de notas comunitarias en sus servicios de la UE, puede hacerlo, pero cualquier sistema de moderación elegido debe ser efectivo para eliminar contenido ilegal conforme al DSA (Euronews, 2025). Esa precisión no era menor, dado que la Comisión ya mantenía una investigación formal sobre X por sus incumplimientos del DSA en materia de moderación basada en usuarios, el mismo modelo que Meta adoptaba.

En el Parlamento Europeo, los eurodiputados Nicolae Ștefănuță y Alexandra Geese presentaron el 14 de enero de 2025 la pregunta parlamentaria escrita E-000110/2025, en la que interrogaron a la Comisión sobre el estado de aplicación del Código de Prácticas contra la Desinformación y sobre si Meta había informado a las autoridades europeas de los cambios previstos en sus políticas de moderación (Parlamento Europeo, 2025). El 20 de enero de 2025, la Comisión integró formalmente el Código de Conducta revisado contra el discurso de odio ilegal en el marco jurídico del DSA (Comisión Europea, 2025a). El desenlace de esa tensión regulatoria se materializó en octubre de 2025, cuando la Comisión emitió conclusiones preliminares determinando que Meta incumple las obligaciones de transparencia del DSA, al imponer mecanismos de denuncia de contenidos ilegales excesivamente complejos y al no ofrecer herramientas adecuadas de apelación a los usuarios de la UE, con la posibilidad de multas de hasta el 6% de los ingresos globales anuales (Comisión Europea, 2025b).

En el ámbito legislativo estadounidense, el Subcomité de Crimen y Contraterrorismo del Comité Judicial del Senado celebró el 9 de abril de 2025 la audiencia "A Time for Truth: Oversight of Meta's Foreign Relations and Representations to the United States Congress", bajo la presidencia del senador Josh Hawley y con la participación del senador Dick Durbin, que incluyó testimonio sobre la responsabilidad corporativa de Meta en sus prácticas de moderación y sus efectos sobre la seguridad de los usuarios (Comité Judicial del Senado, 2025). En conjunto, estas reacciones evidencian que la estrategia de desregulación de Meta no opera en un vacío político, sino que pone en marcha mecanismos de control supranacional cuya efectividad está aún por consolidarse.

El fortalecimiento de una soberanía digital corporativa transforma radicalmente la estructura de la esfera pública e instala a las grandes plataformas en el centro de la configuración política y social contemporánea, permitiéndoles intervenir activamente en la construcción de realidades según intereses privados. Bajo una retórica engañosa que equipara la desregulación con la libertad de expresión y la autorregulación con transparencia, estas corporaciones tecnológicas instauran mecanismos opacos que evaden tanto el control democrático como las limitaciones estatales tradicionales. El desafío central radica en revelar cómo estas nuevas formas de gobernanza digital han dado lugar a un régimen de control informacional que amenaza directamente los fundamentos mismos del debate democrático, subordinando el interés público a lógicas corporativas que operan más allá de toda rendición de cuentas.

Los modelos de moderación de contenido

La arquitectura operativa de Meta se apoyaba en un sistema híbrido de inteligencia artificial (IA), revisión humana y alianzas externas, articulado mediante modelos de aprendizaje automático y procesamiento de lenguaje natural (notoriamente BERT) orientados a detectar violaciones a sus políticas de contenido (discurso de odio, desinformación, violencia gráfica). Pese al entrenamiento con millones de datos etiquetados, la ambigüedad semántica y la variabilidad contextual generaban fallas sistemáticas (falsos positivos y negativos), lo que obligaba a la intervención de revisores humanos y verificadores externos, incrementando costos y limitando la participación.

Meta delegaba parte de su función moderadora a verificadores acreditados por la International Fact-Checking Network (IFCN), quienes aplicaban protocolos definidos en gran medida por las propias plataformas. Según Lauer y Graves (2024), esta tercerización apuntaba a un esquema de gobernanza informativa orientado al interés público en contextos transnacionales. Las medidas adoptadas incluían etiquetas de advertencia, reducción algorítmica del alcance y eliminación del contenido.

La moderación externa opera, no obstante, en condiciones estructuralmente opacas. Gillespie (2022) advierte sobre el alto grado de discrecionalidad en la aplicación de mecanismos de reducción de visibilidad, sin garantías procesales ni transparencia sobre sus parámetros operativos. Las plataformas suelen resistirse a un escrutinio público de estas prácticas, que, pese a no ser del todo desconocidas, se ocultan bajo una ambigüedad funcional que obstaculiza su fiscalización. Los retrocesos en las alianzas entre Meta y los verificadores de la IFCN afectaron directamente la sostenibilidad financiera de actores periodísticos y debilitaron los marcos de supervisión externa, profundizando el desequilibrio entre autorregulación corporativa y exigencias de accountability. Esta reconfiguración otorga mayor autonomía de decisión a la empresa, erosionando los mecanismos de control sobre sus prácticas de gobernanza informativa.

Simultáneamente, la restricción de la moderación a infracciones manifiestas ignora la progresividad de fenómenos como el discurso de odio y la violencia simbólica. La implementación de una política permisiva, bajo el argumento de proteger la libertad de expresión, omite el carácter estructural de las violencias digitales, generando efectos desproporcionados sobre grupos vulnerabilizados. En ausencia de moderación proactiva, se consolida un entorno de hostilidad discursiva que distorsiona el debate público, limita el pluralismo informativo y excluye voces subalternizadas del espacio deliberativo digital.

La postura de Meta, centrada en reducir costos y responsabilidades en materia de moderación bajo el pretexto de defender la libertad de expresión, refleja una tendencia de ciertos gigantes tecnológicos a debilitar los filtros de verificación y

suprimir funciones moderadoras esenciales. Diversos analistas advierten que esta flexibilización pone en riesgo la integridad del debate público y la protección de derechos humanos de millones de usuarios, generando un conflicto geopolítico en el que la soberanía estatal, particularmente la de países latinoamericanos, se ve amenazada por las directrices de las empresas; el propio Zuckerberg ha enfatizado que la plataforma solo bloqueará “contenidos ilegales” y “violaciones graves”, tras admitir la eliminación indebida de publicaciones legítimas (Kaplan, 2025). Conforme al manual de ARTICLE 19 (2023), la moderación de contenidos requiere un conjunto de medidas y herramientas para enfrentar material ilícito o contrario a las normas comunitarias, pero a menudo carece de transparencia y fundamentos robustos. La UNESCO (2023) ha remarcado la necesidad de mayor apertura en las prácticas de las plataformas, sobre todo en América Latina, donde las discusiones sobre derechos digitales avanzan con particular intensidad. En línea con ello, ARTICLE 19 (2023) señala que los algoritmos pueden primar la búsqueda de lucro sobre la seguridad de los usuarios, fomentando la circulación de contenido dañino.

El caso de Brasil ilustra los dilemas derivados de la flexibilización de políticas de Meta. Pese a su posible contradicción con la Constitución Federal y normas antidiscriminatorias, la compañía implementó la nueva “Política de Conducta de Odio” y, tras notificaciones de la Procuraduría Nacional de Defensa de la Democracia, confirmó la rescisión de contratos con verificadores externos (restringida, de momento, a Estados Unidos). Las autoridades brasileñas advirtieron que dichas directrices podrían “generar un terreno fértil para la violación” de derechos fundamentales (Procuradoria Nacional de Defesa da Democracia [PNDD], 2025). El ataque armado en el asentamiento Olga Benário de Tremembé (São Paulo), celebrado sin repercusión alguna en las plataformas de Meta, reforzó la percepción de que el relajamiento de la moderación puede facilitar apologías de la violencia (Lefèvre, 2025). El 20 de enero, la Advocacia-Geral da União (AGU) y varios ministerios convocaron a una audiencia pública para analizar la conformidad de estas políticas con el ordenamiento local (Lefèvre, 2025).

En el otro extremo, las políticas autorregulatorias de países como México o Argentina han dejado que estos cambios sean implementados sin muchos reparos. En México, la Red en Defensa de los Derechos Digitales (R3D, 2025) advirtió que las modificaciones a la política de Hateful Conduct habilitarían discursos que califican a personas trans, mujeres y migrantes como anormales o enfermas mentales, y que la sustitución de los verificadores externos por notas comunitarias elimina un mecanismo que, pese a sus limitaciones, actuaba como barrera frente a la desinformación en su fase de mayor viralización. La preocupación resulta particularmente aguda en un contexto donde las amenazas digitales contra periodistas, activistas y comunidades LGBTIQ+ constituyen una práctica documentada, y donde la moderación proactiva de Meta había funcionado, de manera imperfecta, como contrapeso a dinámicas de acoso sistemático. La organización

ARTICLE 19 también señaló que el momento y el tono del anuncio sugerían que Meta respondía a intereses políticos específicos antes que a una evaluación genuina del impacto de sus políticas sobre la libertad de expresión (ARTICLE 19, 2025).

En Argentina, la Asociación por los Derechos Civiles (ADC), que desde 2020 ha documentado de manera sistemática el vínculo entre la moderación de contenidos de las plataformas y la proliferación de discursos de odio contra comunidades lésbicas, trans y migrantes en Argentina y la región, ha destacado que los modelos de moderación basados en revisión algorítmica presentan fallas estructurales vinculadas a diferencias idiomáticas y contextos socioculturales que las plataformas tienden a ignorar, lo que afecta de manera desproporcionada a grupos ya marginados (ADC, 2020). En ese marco, la decisión de Meta de reducir la moderación proactiva a infracciones de alta gravedad y transferir al sistema de denuncias de usuarios la detección de discursos de odio de menor escala implica que contenidos que equiparan la identidad trans o la homosexualidad con enfermedades mentales, explícitamente permitidos por las nuevas normas comunitarias, circularán sin intervención de la plataforma a menos que un volumen suficiente de usuarios los reporte. Este retroceso en la moderación adquiere una dimensión particular en un país donde la Ley de Identidad de Género (Ley 26.743) y la Ley de Matrimonio Igualitario constituyeron avances normativos de referencia regional, y donde la permisividad de Meta hacia esos discursos entra en tensión directa con el marco de protección antidiscriminatoria vigente, profundizando la brecha entre las obligaciones jurídicas del Estado y la autonomía con la que opera la plataforma dentro de su territorio.

En este escenario, el 'giro trumpista' de Zuckerberg intensifica la disputa sobre soberanía digital y gobernanza de las plataformas, toda vez que combina la resistencia activa de las big techs frente a los marcos regulatorios e instituciones judiciales de ciertos países con la ausencia de regulación efectiva en otros. La profesora Daphne Keller, de Stanford (Keller, como se cita en Lefèvre, 2025), alerta que la reducción o simplificación de los sistemas de detección automática de discurso de odio puede propiciar “una ola real de abusos” contra grupos vulnerables. La ampliación de la jurisdicción del Oversight Board de Meta a la plataforma Threads no disipa las dudas, dadas la lentitud y la naturaleza no siempre vinculante de sus pronunciamientos. Se vislumbra, de este modo, un choque entre la autorregulación corporativa y la autoridad de los Estados, con repercusiones globales en la tutela de derechos fundamentales y en la determinación de los límites del discurso público en la era digital.

Organizaciones de la sociedad civil denuncian que las nuevas directrices de Meta constituyen un retroceso en la protección de derechos fundamentales en entornos digitales, especialmente para comunidades racializadas, mujeres, población LGBTQ+ y minorías religiosas (R3D, 2025). Esa situación demanda una regulación integral que vaya más allá de los controles de competencia e incorpore mecanismos

efectivos de rendición de cuentas y protección de datos personales, en consonancia con los estándares internacionales de derechos humanos y el marco de la OCDE sobre gobernanza digital.

La problemática trasciende la moderación de contenidos y evidencia la necesidad de transparencia en los criterios algorítmicos de visibilidad, así como la responsabilidad de las plataformas ante campañas de acoso digital (Hartmann et al., 2022; IRISBH, 2025). La falta de supervisión externa y auditoría independiente perpetúa un modelo de gobernanza opaco que vulnera la libertad de expresión, la no discriminación y la dignidad de las personas. Según Fernanda Rodrigues (IRISBH, 2025), las directrices de Meta, que mantienen la automatización solo en casos extremados (terrorismo, explotación infantil), transfieren al usuario la carga de la moderación y toleran excesivamente discursos de odio y desinformación.

La permisividad, ilustrada por la equiparación de la homosexualidad o la transexualidad con supuestas “enfermedades mentales”, contraviene disposiciones de organismos internacionales de salud y derechos humanos (OMS, CIE-11; Consejo Federal de Psicología de Brasil) y agrava la posibilidad de viralización de contenidos discriminatorios. Desde el enfoque jurídico, estas políticas colisionan con tratados como la Convención Interamericana contra toda Forma de Discriminación e Intolerancia y legislación nacional antidiscriminatoria en países como Brasil, Argentina y México. La externalización de la moderación mediante sistemas de denuncia y la falta de uniformidad normativa deterioran la salvaguardia de los derechos fundamentales y eximen a las plataformas de asumir plena responsabilidad por el daño que generan sus políticas permisivas.

Aunque Meta argumenta que este régimen reducirá la supresión de contenido legítimo e incrementará el número de revisores humanos, persiste la inquietud sobre la coherencia e imparcialidad de sus decisiones, dada la falta de transparencia en la aplicación de las normas. Además, la escasa adecuación a la normativa local podría derivar en litigios, suspensiones y un agravamiento del conflicto geopolítico entre Estados que reivindican su soberanía regulatoria y las grandes empresas tecnológicas (IRISBH, 2025).

En un contexto en el que la viralidad permite que narrativas sin sustento se difundan con facilidad, la autorregulación corporativa, ya cuestionada por su ineficacia, se muestra todavía más limitada, pues no todos los usuarios disponen de la formación o el tiempo necesario para analizar de manera rigurosa la veracidad de los contenidos que consumen y comparten. Este escenario se encuentra vinculado con el discurso del presidente Donald Trump y de sectores conservadores en Estados Unidos, que consideran “censura” toda moderación de contenidos ofensivos o falsos. La situación adquiere matices geopolíticos cuando se disputa la legitimidad de los gobiernos y tribunales de América Latina o Europa para regular a las big tech,

previéndose un incremento de la judicialización del tema y un recrudecimiento del debate sobre la responsabilidad de las corporaciones digitales en la proliferación de desinformación y discursos de odio. En Brasil, por ejemplo, el Supremo Tribunal Federal y el Congreso Nacional discuten propuestas de reforma y proyectos de ley, como el PL 2630/2020, dirigidos a la regulación de las redes sociales.

El año 2024 estuvo marcado por un intenso debate regulatorio orientado a delimitar la responsabilidad objetiva y subjetiva de las plataformas digitales en materia de moderación de contenidos. El 2025 prometía avanzar hacia un modelo normativo más equilibrado y garantista, que armonizara la protección de la libertad de expresión con la mitigación de los riesgos inherentes a la desinformación, el discurso de odio y otras formas de violencia digital. No obstante, los recientes anuncios de Meta parecen reconfigurar estas perspectivas, planteando interrogantes políticos y geopolíticos de amplio alcance. La reivindicación de una libertad de expresión prácticamente irrestricta conlleva el riesgo de legitimar prácticas contrarias a la dignidad humana, la igualdad y la seguridad digital. Dicho enfoque, próximo a un *laissez-faire* en la gobernanza de plataformas, se traduce en la atenuación de las obligaciones de moderación y en un retroceso en la protección contra la discriminación y la violencia simbólica. Esta situación no se circunscribe al territorio brasileño y adquiere una dimensión transnacional y afecta a cualquier jurisdicción que pretenda ejercer control soberano sobre las grandes corporaciones tecnológicas.

Dentro de este panorama, se intensifica la confrontación entre los regímenes de autorregulación corporativa y los marcos normativos diseñados en diálogos con la sociedad civil y los Estados para preservar la vigencia efectiva de los derechos fundamentales en entornos digitales. Tal pugna pone en cuestión la capacidad de los Estados para regular el espacio virtual de manera soberana y reactiva la necesidad de articular estándares internacionales vinculantes para la gobernanza de plataformas. Ello se considera esencial a fin de impedir que la discrecionalidad de las grandes compañías defina unilateralmente los límites del discurso público en la era digital, con las consecuentes implicaciones para la tutela de la libertad de expresión, la no discriminación y el acceso a la información veraz.

Cambio en las normas comunitarias y el fin de los fact-checkers: las nuevas reglas de Meta

A la par de los cambios en la política de moderación de contenidos a través de la eliminación de los verificadores externos, Meta actualizó su política sobre contenidos de odio, denominada ahora “Hateful Conduct” (Meta, 2025a). La compañía emprendió una serie de transformaciones en torno a las reglas que determinan qué tipo de contenidos están permitidos y cuáles no, más precisamente, en las normas comunitarias sobre las que se sustentan los discursos que circulan en las plataformas

de Meta. Estas normas se constituyen como el gran paraguas sobre el que se definen las reglas básicas de convivencia en la plataforma, delimitando el tipo de contenido que no es aceptado y aquel que requiere cierto contexto y/o mensaje de advertencia a su público potencial, pues la plataforma asume un papel de censor en pos de evitar la propagación de discursos discriminatorios, segregacionales y violentos que vulneran los derechos humanos.

El papel que cumplen las normas comunitarias es fundamental para delimitar la labor de moderación de contenidos que realiza la propia plataforma: la moderación gira en torno al respeto de esas normas comunitarias, de modo que cualquier cambio en ellas modifica el acto mismo de la moderación. Tanto los procesos automatizados, que alcanzan más del 90% de las acciones moderadoras, como la realizada por humanos, se rigen bajo estas normas comunitarias, las cuales fungen como el marco sobre el que se decidirá la suerte de un contenido. Si bien las normas comunitarias son admitidas por los usuarios a la hora de aceptar los términos y condiciones de la plataforma, se considera que son decisiones unilaterales que toma la compañía y ante las cuales los usuarios están obligados a aceptarlas o, por el contrario, abandonar la red social, evidenciando así la dinámica autorreguladora y discrecional en la cual el poder se centraliza a tal nivel de no dejar otras alternativas a sus millones de usuarios. Esta verticalidad operativa determina tanto la aceptación de dichas condiciones en un determinado momento como la aceptación implícita de los futuros cambios que desde la plataforma se lleven adelante, bajo la lógica dualista de admitirlas o abandonar la red.

Bajo el pretexto de ampliar las libertades de los usuarios en el ejercicio de la libertad de expresión, Mark Zuckerberg decidió modificar aspectos sustanciales de las normas comunitarias vinculadas a limitar la propagación de discursos de odio. Desde el 7 de enero de 2025, esas modificaciones se centran en el inciso Conducta que incita al odio (Meta, 2025a), particularmente en relación a temas de género y migración. Aunque se declara que “no permitimos la conducta de odio en Facebook, Instagram y Threads”, dichas modificaciones se presentan como un cambio de paradigma en la política de contenidos de la plataforma. Entre las actualizaciones más relevantes se cuentan los siguientes cambios en las normas sobre discursos de odio.

- La supresión del término *discurso* por el de *conducta* que incita al odio. En este caso, el cambio conceptual adquiere un carácter de mayor imprecisión, acompañado de una falta de delimitación de lo que es una conducta en el entorno digital.
- Se eliminan, asimismo, una serie de principios que buscaban salvaguardar la integridad de las personas cuando son afectadas o atacadas por sus *características protegidas* (raza, etnia, nacionalidad, discapacidad, afiliación religiosa, orientación de género e identidad sexual, sexo y enfermedad grave), quedando permitidas aquellas “*declaraciones de inferioridad, expresiones de*

desprecio o disgusto; maldiciones; y llamados a la exclusión o segregación”, como así también el discurso violento sustentado en estas características protegidas de las personas.

- El cambio en las normas comunitarias en torno a las conductas que incitan al odio también abren la puerta a la posibilidad de generar contenidos dirigidos a una persona o grupo de personas sobre la base de sus características protegidas antes mencionadas o su estatus migratorio sin ninguna restricción, haciendo foco en la equiparación con objetos, estados no humanos o enfermedades; la consideración de mujeres como objetos domésticos y o la identificación con algún tipo de enfermedad, ya sea cáncer, enfermedades de transmisión sexual, entre otras.
- Se permiten las conductas que tengan como objetivo la exclusión de espacios en materia de género, particularmente *“como baños, ligas deportivas, grupos de salud y apoyo y escuelas específicas”*.
- El cambio en las normas comunitarias abre paso a los discursos de *“inferioridades enfocadas en apariencias físicas incluyendo, pero no limitado a: feo, horrible”*, y en características mentales como por ejemplo: tontos, estúpidos, idiotas, enfermo mental, retrasado, loco, entre otros. Desde esta perspectiva, las nuevas normas comunitarias permiten señalamientos de enfermedad mental o anormalidad cuando se basa en el género o la orientación sexual, derivado de un discurso político y/o religioso sobre el transgénero y la homosexualidad.

Las actualizaciones de las normas comunitarias con respecto a las ahora conductas de odio tienden a constituir un claro retroceso en materia de moderación de contenidos, particularmente en lo que se refiere a las características protegidas de las personas. Bajo el argumento de que las reglas eran “demasiado restrictivas y propensas a una aplicación excesiva” (Meta, 2025b) y que se estaba limitando el debate político legítimo a través de la censura, la plataforma pone sobre la mesa la carta de fortalecer la libertad de expresión, pero a costa de la vulneración de derechos humanos de los usuarios a través de la ampliación de discursos de odio sobre temáticas de género y migración. El resultado contradice esa premisa. Estas actualizaciones abren la puerta a un sinfín de discursos que tienen el único objetivo de erosionar el debate público a través de una serie de descalificaciones, discriminación y segregación con una fuerte repercusión en el plano material de las relaciones sociales. Al permitir asemejar a las personas con objetos o enfermedades de todo tipo, se profundiza la degradación de la esencia humana, al llevar un intercambio de ideas a un plano con fuerte carga de violencia simbólica. Las actualizaciones de las normas comunitarias habilitan, además, un camino hacia la segregación por la condición de género de determinados espacios públicos, con la deriva de establecer qué puede y qué no puede hacer una persona y la tolerancia que permite la plataforma hacia ciertos discursos políticos y/o religiosos donde se señale

como enfermos mentales a personas por su orientación sexual, marca un panorama en extremo preocupante en materia de derechos humanos.

Aunque en otros tiempos “La creación y modificación de las normas se basan en los comentarios que reciben de la comunidad de personas usuarias y en los consejos de especialistas en campos como tecnología, seguridad pública y derechos humanos” (ARTICLE 19, 2020), estos últimos cambios parecen responder a otros intereses. Tanto la política migratoria como la agenda de género se han constituido como dos de los pilares de la política del mandatario estadounidense Donald Trump, frente a los cuales, los directivos de las grandes corporaciones tecnológicas han decidido poner las plataformas a su disposición, a través de la relajación de estas normas comunitarias. Si bien estos cambios se irán implementando paulatinamente en todo el mundo, se espera que en el ámbito de la Unión Europea tengan mayor dificultad de hacerlo, al considerar el marco normativo en la región, política regulatoria que Zuckerberg considera como “institucionalización de la censura por su creciente cantidad de leyes”.

Otro de los cambios más radicales que presentó la compañía, es la eliminación de los mecanismos de verificación de contenidos por entidades externas a Meta, los fact-checkers. La labor de estas organizaciones periodísticas comenzó en 2016 bajo el principio de brindar información a los usuarios sobre la veracidad o falsedad de cierto contenido al que estaban expuestos. Si bien estas entidades nunca estuvieron facultadas para determinar la eliminación de contenidos de la plataforma, su principal aporte era brindar información que contextualizara a los usuarios y emitir una suerte de alerta sobre el carácter verídico (o no) de dichos contenidos.

La labor de los verificadores externos comenzaba en dos circunstancias. La primera era cuando los mecanismos automatizados de moderación de la plataforma “identifican historias potencialmente falsas” [traducción de los autores] (Lyons, 2018) principalmente cuando se detectan comentarios de usuarios aludiendo la falsedad sobre un contenido y deciden derivar el caso a estas organizaciones. También de forma proactiva cuando los verificadores detectan un caso problemático que requiere su intervención. Posteriormente, los verificadores evaluaban el caso, “califican su exactitud y escriben un artículo explicando los hechos detrás de su calificación” [traducción de los autores] (Lyons, 2018). Ante la detección de la desinformación y posterior evaluación sobre el impacto que puede llegar a tener su difusión, la plataforma reduce su distribución en su News Feed y emite un mensaje de alerta a aquellos usuarios que quieran compartir el contenido en cuestión, brindando la información contextual desarrollada por los verificadores.

Este mecanismo generó una suerte de expectativa en la política moderadora de la plataforma que, en conjunto con la creación del Oversight Board, abría la puerta a que actores externos pudieran intervenir, en cierta medida, en las

orientaciones sobre la moderación de conductas de los usuarios de la plataforma. El contexto no es casual. Escándalos como la filtración masiva de datos y las campañas de desinformación esgrimidas a través de la plataforma, llevaron a una apertura que les permitiera evidenciar cambios profundos en su política interna; ese esquema llegó a su fin el pasado 7 de enero de 2025 para dar lugar al programa de Community Notes que Meta comenzó a implementar en Estados Unidos a partir de ese año, con una expansión progresiva prevista hacia el resto del mundo. Esta transición implica un cambio de paradigma en la moderación de contenidos que evidencia el desplazamiento operativo de un grupo reducido de expertos, los verificadores, hacia los propios usuarios de la plataforma como agentes moderadores. El cambio de rumbo hacia un modelo similar al de la plataforma X se presenta como un esquema horizontal en la que los usuarios informan a otros sobre publicaciones potencialmente engañosas, una suerte de crowdsourcing³ colaborativo en el cual todos los usuarios son potenciales moderadores de contenidos. Los estudios realizados sobre este mecanismo en la red X señalan, no obstante, que no hay evidencia de que la introducción de este tipo de moderación sea eficaz para reducir el compromiso con la desinformación en las redes sociales. “Nuestros hallazgos sugieren que las notas comunitarias podrían ser demasiado lentas para reducir eficazmente el compromiso con la desinformación en la fase de difusión temprana (y más viral)” (Chuai et al., 2024, p. 1).

Si bien ambos modelos de moderación pueden presentar sus puntos fuertes y débiles, contraponiendo la noción de participación de los usuarios con la de opiniones de expertos en la materia, lo cierto es que esta transición se da en un marco de retirada de una serie de compromisos para promover un espacio de discusión plural, democrática y libre de violencia, bajo el discurso de promover mayores libertades en torno a la libertad de expresión. Esa decisión presenta ambos modelos como opuestos, argumentando que no podrían ser complementarios, algo que resulta poco verosímil ya que se podría fortalecer la acción de moderación desde las opiniones y comentarios de los usuarios con las recomendaciones de las diversas instituciones internacionales de verificación de contenidos.

La eliminación de los fact-checkers abre la puerta a una mayor circulación de información no verificada y discursos de odio, y plantea interrogantes sobre el marco de responsabilidades que se aplican a las plataformas, que terminan por ser empresas autorreguladas con la consecuencia de no dejar en claro quién asumirá la rendición de cuentas ante la eventual propagación de información falsa o contenido perjudicial. La empresa conserva su control sobre los algoritmos que determinan la visibilidad de las publicaciones y sigue capitalizando económicamente la interacción generada por la viralización del contenido, pero, al mismo tiempo, diluye su responsabilidad al redefinir los límites de su intervención. Este nuevo paradigma configura un escenario donde las plataformas ejercen menos poder de moderación en términos normativos, desligándose de las responsabilidades frente a los usuarios, pero mantienen intacto

su dominio estructural sobre la información que circula, sin asumir las consecuencias que esta nueva dinámica puede desencadenar.

El lugar del Oversight Board ante los cambios de Meta

Los cambios radicales en las políticas de contenidos mencionados pusieron en jaque otro gran estandarte en materia de moderación de Meta como lo es el Oversight Board, poniendo un grado de incertidumbre sobre su labor supervisora y el futuro como institución independiente a nivel operativo, pero que se desempeña gracias a los recursos económicos que la plataforma dispone. El Oversight Board se constituyó en 2020 como una nueva instancia de apelación para los usuarios de la plataforma a la que pueden recurrir en el caso de que las decisiones moderadoras, tanto automatizadas como mediadas por humanos de Meta consideren que no son pertinentes o adecuadas en relación a las resoluciones de retirar o mantener en línea un determinado contenido. Al agotar estas instancias, el usuario puede solicitar que su caso sea revisado por el Oversight Board, quien los seleccionará bajo criterios de dificultad (con relación a las normas de moderación de la plataforma) e importancia (con relación al impacto social en materia de libertad de expresión, privacidad, etc.) del caso con la finalidad de generar cierta jurisprudencia en las políticas moderadoras de la compañía.

El Oversight Board actúa desde dos ángulos complementarios. Por un lado, toma una decisión que es vinculante e inapelable para la plataforma sobre un caso en particular que haya seleccionado. Esto implica, por ejemplo, reponer determinado contenido que, a su criterio, haya sido eliminado de forma incorrecta o determinar la eliminación de un contenido que viole las normas comunitarias, en el caso que la plataforma haya decidido sostener una publicación. Las decisiones se construyen bajo los parámetros de la propia reglamentación de la plataforma, es decir, según lo que señalan las normas comunitarias, las políticas de contenidos de Meta y sus valores (Oversight Board, 2019) y, por último, en normas del derecho internacional, particularmente en los “Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos (PRNU)”, los cuales, es importante aclarar, guardan “un marco voluntario para las responsabilidades de las empresas privadas en materia de derechos humanos” (PRNU, 2011), un marco que carece de carácter vinculante. En este punto, el “marco legal” bajo el que actuará el Oversight Board son las normas comunitarias, las cuales, como se ha mencionado, se han modificado en su inciso “Conducta que incita al odio”. Este cambio ata de manos al Oversight Board ya que sus decisiones vinculantes en relación con preservar o eliminar un contenido se deben justificar en dichas normas, por lo que no podrá intervenir o contrariar a la plataforma en casos donde discursos de odio sean, ahora, permitidos.

Otra de las facultades del Oversight Board se enfoca en la emisión de recomendaciones a la plataforma de carácter no vinculante en relación al diseño de las políticas de contenido y normas comunitarias. Las recomendaciones emitidas tienen que partir de casos en particular en los cuales, por ejemplo, se promueva el discurso de odio por cuestiones de género o a las personas migrantes; el Oversight Board podría emitir una serie de recomendaciones sobre la necesidad de fortalecer las normas comunitarias en materia de discursos que incitan al odio.

Considerando los recientes cambios de Meta y las perspectivas políticas a mediano plazo, esas recomendaciones no tendrían una respuesta positiva. También es importante mencionar que la plataforma puede solicitarle al Oversight Board recomendaciones sobre determinados casos o políticas a implementar. Un ejemplo de esto fue cuando Facebook decide suspender indefinidamente la cuenta del actual mandatario de EE. UU., Donald Trump, recurriendo al Oversight Board para que emita su opinión sobre dicha acción. No obstante, las decisiones tomadas el 7 de enero de 2025, la plataforma no recurrió a la opinión o consulta de los expertos que conforman el Oversight Board, algo que hubiese sido de gran valor al poner sobre la mesa un contrapunto ante la modificación de las políticas de moderación.

Frente a la eliminación de los verificadores externos, el Oversight Board se presenta como la única institución independiente de la plataforma en el ámbito de la moderación de contenidos, pero con grandes limitaciones en la materia, tanto desde lo esgrimido en su estatuto fundacional como así también por los cambios acaecidos a comienzos del 2025. Esto evidencia tanto la creciente concentración corporativa de las decisiones moderadoras como la fragilidad con la que fueron constituidas estas instituciones, quedando a merced de los vientos políticos e ideológicos que asumen los gerentes de la compañía.⁴

Conclusiones

Las recientes transformaciones en las políticas de moderación de Meta, justificadas bajo la retórica de la "libertad de expresión" y la supuesta necesidad de reducir la censura, han expuesto una tensión estructural entre la autonomía corporativa y la protección de derechos fundamentales en la esfera digital. Al eliminar la verificación independiente de hechos y reducir la moderación automatizada, Meta ha transferido gran parte de la responsabilidad a los usuarios, un enfoque que no garantiza una autorregulación efectiva ni transparente. Por el contrario, este modelo descentralizado facilita la proliferación de discursos de odio, desinformación y violencia simbólica, erosionando la calidad del debate público y generando posibles entornos hostiles para comunidades históricamente vulnerabilizadas, como mujeres, personas LGBTIQ+, migrantes y minorías raciales.

El contexto político actual, marcado por la alineación entre las grandes corporaciones tecnológicas (big techs) y el gobierno de Donald Trump, refuerza la necesidad de repensar la soberanía digital como un espacio de tensión regulatoria. La decisión de Meta de modificar sus normas comunitarias, especialmente en lo referente a la "conducta que incita al odio", reconfigura la relación de la plataforma con los marcos jurídicos de múltiples jurisdicciones (como se evidencia en el caso brasileño) y pone de manifiesto la asimetría de poder entre las empresas transnacionales y los Estados. Este escenario revela la existencia de vacíos normativos y la urgencia de establecer marcos legales coherentes y con vocación global, capaces de abordar los desafíos que plantea la gobernanza digital en un mundo cada vez más interconectado.

La creación de instancias como el Oversight Board prometía mayor independencia y legitimidad en la resolución de disputas sobre moderación de contenido; su capacidad de incidir en las políticas de Meta está limitada por la naturaleza unilateral de las decisiones corporativas, como lo demuestra la reciente modificación de las normas comunitarias sin consultar a este órgano. La desvinculación de los verificadores externos afiliados a redes como la International Fact-Checking Network (IFCN) ha abierto una brecha en la verificación de contenidos, potenciando la circulación de información engañosa y desinformación. Esta dinámica dificulta que los mecanismos judiciales o administrativos existentes actúen con oportunidad y eficacia, lo cual tensiona aún más la relación entre los sistemas de autorregulación empresarial y la intervención estatal, especialmente ante posibles violaciones de derechos humanos y normas constitucionales.

La experiencia de Meta ilustra una tendencia global preocupante. La búsqueda de "flexibilización" en la moderación digital coincide con el aumento de la polarización política y la retórica de la libre expresión irrestricta. Esta supuesta desregulación invisibiliza la responsabilidad de las plataformas en la curaduría del discurso y debilita la tutela de valores esenciales, como la dignidad humana, la no discriminación y la igualdad. Ante este panorama, se hace imprescindible reforzar la transparencia algorítmica, promover la rendición de cuentas en cada nivel de la cadena de moderación y diseñar mecanismos de supervisión externos dotados de legitimidad y eficacia. Solo así se podrá garantizar que las plataformas digitales no se conviertan en espacios de impunidad para discursos nocivos y prácticas discriminatorias.

La situación actual de Meta redefine los equilibrios entre libertad y seguridad en el espacio digital y proyecta un reto mayor a la gobernanza digital: garantizar que la innovación tecnológica y las dinámicas del mercado global no vulneren los cimientos democráticos y las garantías fundamentales de los usuarios. Abordar ese reto exige una articulación multiactor que involucre al sector público, la sociedad civil, organismos internacionales y las propias corporaciones, poniendo en el centro la protección de los derechos humanos y la responsabilidad social de las empresas. En ese marco, cabe formular un conjunto de recomendaciones concretas. Los

organismos reguladores nacionales y supranacionales deben exigir evaluaciones de riesgo previas a cualquier modificación de los modelos de moderación en plataformas de alcance sistémico, adoptando el estándar del DSA europeo como referencia mínima. Los Estados de América Latina deben avanzar en marcos normativos propios que reconozcan la responsabilidad objetiva de las plataformas ante daños documentados por discursos de odio. La sociedad civil y la academia deben impulsar auditorías algorítmicas independientes con criterios de derechos humanos. Las propias plataformas deben restablecer mecanismos de verificación externa y garantizar que los sistemas de apelación de decisiones de moderación sean accesibles, transparentes y vinculantes. La gobernanza digital del siglo XXI requiere que el poder y las ganancias corporativas estén acompañados de responsabilidades concretas, exigibles y urgentes.

Referencias

- ARTICLE 19. (2020). *Sobre las Normas Comunitarias de Facebook*. https://seguridadintegral.articulo19.org/wp-content/uploads/2020/10/Article19_ManualFB_Web.pdf
- ARTICLE 19. (2023). *Manual sobre moderación de contenido y libertad de expresión en plataformas digitales*. https://www.article19.org/wp-content/uploads/2024/03/FINAL_SPANISH-v.4.-SM4P-Content-moderation-handbook-.pdf
- ARTICLE 19. (2025, enero). *Meta: Prioritise human rights, not politics*. ARTICLE 19. <https://www.article19.org/resources/meta-prioritise-human-rights-not-politics/>
- Asociación por los Derechos Civiles (ADC). (2020, 1 de octubre). *Discurso de odio y moderación de contenidos*. <https://adc.org.ar/2020/10/01/discurso-de-odio-y-moderacion-de-contenidos/>
- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), 1-52.
- Comisión Europea. (2025a, 20 de enero). *Código de Conducta revisado sobre discurso de odio ilegal integrado en el marco del DSA*. Comisión Europea. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en
- Comisión Europea. (2025b, 24 de octubre). *Conclusiones preliminares: Meta incumple obligaciones de transparencia del DSA*. Comisión Europea. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503
- Comité Judicial del Senado de los Estados Unidos. (2025, 9 de abril). *A Time for Truth: Oversight of Meta's Foreign Relations and Representations to the United States Congress*. Subcomité sobre Crimen y Contraterrorismo. <https://www.judiciary.senate.gov/committee-activity/hearings/a-time-for-truth-oversight-of-metas-foreign-relations-and-representations-to-the-united-states-congress>
- Couture, S., & Toupin, S. (2019). What does the notion of "sovereignty" mean when referring to the digital? *New Media & Society*, 21(10), 2305-2322. <https://doi.org/10.1177/1461444819865984>

- Direitos na Rede. (2025, 8 de enero). *Contra el retroceso en la moderación de contenido de Meta y los ataques a la regulación democrática del espacio digital*. <https://direitosnarede.org.br/2025/01/08/contra-el-retroceso-en-la-moderacion-de-contenido-de-meta-y-los-ataques-a-la-regulacion-democratica-del-espacio-digital/>
- Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, 136(2), 526-607.
- Euronews. (2025, 10 de enero). *Do Meta's announcements run counter to the European regulation on digital services?*. <https://www.euronews.com/next/2025/01/10/do-metas-announcements-run-counter-to-the-european-regulation-on-digital-services>
- Flew, T. (2021). *Regulating platforms*. Polity Press.
- Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, 8(3). <https://doi.org/10.1177/20563051221117552>
- Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22(6), 854-871. <https://doi.org/10.1080/1369118X.2019.1573914>
- Gorwa, R. (2024). *The politics of platform regulation. How governments shape online content moderation*. Oxford University Press. <https://doi.org/10.1093/oso/9780197692851.001.0001>
- Habermas, J. (2025). *Un nuevo cambio estructural de la esfera pública y la política deliberativa* (J. C. Velasco Arroyo, trad.). Trotta. (Obra original publicada en 2022)
- Hartmann, I. A., Curzi, Y., Zingales, N., & Almeida, C. (Orgs.). (2022). *Moderação de conteúdo online: Contexto, cenário brasileiro e suas perspectivas regulatórias*. <https://diretorio.fgv.br/conhecimento/moderacao-de-conteudo-online-contexto-cenario-brasileiro-e-suas-perspectivas>
- IRISBH – Instituto de Referência em Internet e Sociedade. (2025, 16 de enero). *Moderação de conteúdo e redes sociais: o que esperar (de velho) no ano novo?* <https://irisbh.com.br/moderacao-de-conteudo-e-redes-sociais-o-que-esperar-de-velho-no-ano-novo/>
- Jiang, M., & Belli, L. (Eds.). (2025). *Digital sovereignty in the BRICS countries. How the Global South and emerging power alliances are reshaping digital governance*. Cambridge University Press. <https://doi.org/10.1017/9781009531085>
- Kaplan, J. (2025, 7 de enero). *More Speech and Fewer Mistakes*. Meta Newsroom. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- Klonick, K. (2018). The new governors. The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6), 1598-1670.
- Lauer, L., & Graves, L. (2024). How to grow a transnational field: A network analysis of the global fact-checking movement. *New Media & Society*. Advance online publication. <https://doi.org/10.1177/14614448241227856>
- Lefèvre, F. (2025, 15 de enero). *Moderação de conteúdo e plataformas digitais [Columna]*. <https://www.mobiletime.com.br/colunistas/15/01/2025/flavia-lefevre-coluna/>
- Lyons, T. (2018, 14 de junio). *Hard Questions: How Is Facebook's Fact-Checking Program Working?* Meta Newsroom. <https://about.fb.com/news/2018/06/hard-questions-fact-checking/>
- Magalhães, J. C., Iglesias Keller, C., & Gorwa, R. (2026). The great sysop. Elon Musk, X, and the emergence of platform illiberalism. *New Media & Society*. Publicación anticipada en línea. <https://doi.org/10.1177/14614448261424889>
- Meta. (2025a, enero). *Hateful Conduct*. Meta Transparency Center. <https://transparency.meta.com/policies/community-standards/hateful-conduct/>
- Meta. (2025b, 7 de enero). *Más libertad de expresión y menos errores*. Meta Latinoamérica. <https://about.fb.com/ltam/news/2025/01/mas-libertad-de-expresion-y-menos-errores/>

- Naciones Unidas [PRNU]. (2011). *Principios rectores sobre las empresas y los derechos humanos*. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_sp.pdf
- Newton, C. (2026, 2 de abril). *Exclusive: Meta has discussed ending funding to the Oversight Board*. Platformer. <https://www.platformer.news/meta-oversight-board-funding-cancel/>
- Oversight Board. (2019). *Oversight Board Charter*. https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf
- Parlamento Europeo. (2025, 14 de enero). *Pregunta parlamentaria escrita E-000110/2025: Meta and its fact-checking programme*. https://www.europarl.europa.eu/doceo/document/E-10-2025-000110_EN.html
- Pohle, J., & Thiel, T. (2020). Digital sovereignty. *Internet Policy Review*, 9(4). <https://doi.org/10.14763/2020.4.1532>
- Procuradoria Nacional de Defesa da Democracia (PNDD). (2025, 14 de enero). *Notificação extrajudicial à Meta Platforms Inc. sobre políticas de moderação de conteúdo*. Ministério da Justiça e Segurança Pública do Brasil.
- R3D – Red en Defensa de los Derechos Digitales. (2025, 10 de enero). *Cambios en políticas de moderación de contenidos de Meta despiertan alarma por posibles riesgos a la libertad de expresión*. <https://r3d.mx/2025/01/10/cambios-en-politicas-de-moderacion-de-contenidos-de-meta-despiertan-alarma-por-posibles-riesgos-a-la-libertad-de-expresion/>
- Siapera, E., & Viejo-Otero, P. (2021). Governing hate. Facebook and digital racism. *Television & New Media*, 22(2), 112-130. <https://doi.org/10.1177/1527476420982232>
- Suzor, N. P. (2019). *Lawless. The secret rules that govern our digital lives*. Cambridge University Press. <https://doi.org/10.1017/9781108666428>
- UNESCO. (2023, febrero). *Transparencia, moderación de contenidos y libertad de expresión: Perspectivas multi-actor en América Latina*. <https://www.unesco.org/es/articles/transparencia-moderacion-de-contenidos-y-libertad-de-expresion-perspectivas-multi-actor-en-america>

Este artículo es de acceso abierto. Los usuarios pueden leer, descargar, distribuir, imprimir y enlazar al texto completo, siempre y cuando sea sin fines de lucro y se cite la fuente.

CÓMO CITAR ESTE ARTÍCULO:

Souz, Rodrigo Ardisson de y Neubauer, Dardo Emanuel. (2026). Modelos de moderación de contenido, mecanismos de autorregulación y dinámicas de autolegitimación. *Paakat: Revista de Tecnología y Sociedad*, 16(31). <http://dx.doi.org/10.32870/Pk.a16n31.933>

¹ Profesor-Investigador, División de Estudios Jurídicos, CIDE, Ciudad de México. Doctor en Antropología Social, ENAH. rodrigo.ardissom@cide.edu | ORCID: 0009-0002-0640-5052. Agradezco al Programa de Becas Posdoctorales, a la Coordinación de Humanidades de la UNAM y a mi tutor por el apoyo brindado durante mi estancia posdoctoral en el PUEDJS.

² Investigador posdoctoral, PUEDJS-UNAM. Doctor en Ciencias Políticas y Sociales, UNAM. dardoneubauer@politicas.unam.mx | ORCID: 0009-0001-9701-0509. Agradezco al Programa de Becas Posdoctorales, al Secihti y a mi tutor por el apoyo brindado durante mi estancia posdoctoral en el PUEDJS.

³ Colaboración abierta que consiste en externalizar tareas que, tradicionalmente, realizaban miembros de la compañía, empleados o contratistas, dejándolas a cargo de un grupo numeroso de personas o de una comunidad.

⁴ Al cierre de la redacción de esta versión del artículo, la situación del Oversight Board había experimentado un deterioro adicional. En abril de 2026, Meta informó a los miembros del organismo que podría dejar de financiarlo una vez venza su compromiso en 2028, habiendo ya reducido significativamente el presupuesto en 2025 y previendo nuevas reducciones para 2027 y 2028. Ambas partes negocian un acuerdo que podría incluir la creación de una entidad independiente que preste servicios similares a otras plataformas tecnológicas (Newton, 2026). Este proceso de desfinanciamiento confirma y profundiza el diagnóstico aquí formulado sobre la fragilidad estructural de los mecanismos de autorregulación corporativa.