

Justicia privatizada y gobernanza algorítmica: el alcance real del Facebook Oversight Board

Privatized Justice and Algorithmic Governance: The Real Reach of the Facebook Oversight Board

Justiça privatizada e governança algorítmica: o alcance real do Facebook Oversight Board

DARDO EMANUEL NEUBAUER

Profesor en la Facultad de Ciencias Políticas y Sociales (FCPyS) de la Universidad Nacional Autónoma de México (UNAM). Doctor en Ciencias Políticas y Sociales por la UNAM e investigador posdoctoral en el Programa Universitario de Estudios sobre Democracia, Justicia y Sociedad (PUEDJS-UNAM), lugar donde coordina el área jurídica del laboratorio digital Tlatelolco Lab.

Justicia privatizada y gobernanza algorítmica: el alcance real del Facebook Oversight Board¹

Privatized Justice and Algorithmic Governance: The Real Reach of the Facebook Oversight Board

Justiça privatizada e governança algorítmica: o alcance real do Facebook Oversight Board

Dardo Emanuel Neubauer

Facultad de Ciencias Políticas y Sociales, Universidad Nacional Autónoma de México, México

dardoneubauer@politicas.unam.mx (<https://orcid.org/0009-0001-9701-0509>)

Recibido: 26-01-2026 / Aceptado: 07-04-2026

<https://doi.org/10.18800/conexion.202601.002>

RESUMEN

Este artículo analiza el Facebook/Meta Oversight Board como dispositivo de gobernanza privada de la moderación de contenidos y evalúa su incidencia efectiva sobre la moderación automatizada. Ante la opacidad estructural de los sistemas algorítmicos propietarios, el estudio emplea análisis documental cualitativo con lógica comparativa de casos, examinando exclusivamente las recomendaciones del Board y las respuestas de Meta sobre su implementación, a partir de cuatro casos emblemáticos. Los hallazgos muestran una tensión persistente: si bien el Board puede imponer decisiones vinculantes en casos individuales, sus recomendaciones sistémicas, en particular las que buscan afectar el núcleo técnico de la visibilidad

—*ranking*, bancos de coincidencia, *cross-check* y métricas de transparencia—, son aceptadas de forma parcial, pospuestas mediante fórmulas de factibilidad o cerradas sin acciones adicionales. El artículo sostiene que ello consolida un modelo de justicia privatizada, con debido proceso algorítmico debilitado y escasa observabilidad externa.

ABSTRACT

This article examines the Facebook/Meta Oversight Board as a private governance mechanism for content moderation and assesses its effective influence over automated moderation. Given the structural opacity of proprietary algorithmic systems, the study adopts qualitative documentary analysis with a comparative case

¹ Agradezco al Programa de Becas Posdoctorales, a la SECIHTI y a mi tutor por el apoyo brindado en mi estancia posdoctoral en el PUEDJS-UNAM.

logic, focusing exclusively on the Board's recommendations and Meta's implementation responses across four emblematic cases. The findings reveal a persistent tension: while the Board can issue binding decisions in individual cases, its systemic recommendations—especially those aimed at the technical core of visibility governance (ranking, matching banks, cross-check, and transparency metrics)—are implemented only partially, deferred through feasibility language, or closed without further action. The article argues that this configuration consolidates a model of privatized justice, marked by weakened algorithmic due process and limited external observability.

RESUMO

Este artigo analisa o Facebook/Meta Oversight Board como um dispositivo de governança privada da moderação de conteúdos e avalia sua incidência efetiva sobre a moderação automatizada. Diante da opacidade estrutural dos sistemas algorítmicos proprietários, o estudo emprega análise documental qualitativa com lógica comparativa de casos, examinando exclusivamente as recomendações do Board e as respostas de implementação da Meta, a partir de quatro casos emblemáticos. Os achados evidenciam uma tensão persistente: embora o Board possa impor decisões vinculantes em casos individuais, suas recomendações sistêmicas — em particular as que buscam afetar o núcleo técnico da governança da visibilidade

(*ranking*, bancos de correspondência, *cross-check* e métricas de transparência) — são implementadas de forma parcial, adiadas por meio de fórmulas de “viabilidade” ou encerradas sem ações adicionais. O artigo sustenta que essa configuração consolida um modelo de justiça privatizada, com devido processo algorítmico enfraquecido e baixa observabilidade externa.

PALABRAS CLAVE / KEYWORDS / PALAVRAS-CHAVE

Gobernanza algorítmica, moderación de contenidos, Oversight Board, justicia privatizada, debido proceso algorítmico / algorithmic governance, content moderation, Oversight Board, privatized justice, algorithmic due process / governança algorítmica, moderação de conteúdos, Oversight Board, justiça privatizada, devido processo algorítmico

Introducción

La discusión sobre la moderación de contenidos en las plataformas sociodigitales se ha consolidado como uno de los ejes centrales del debate contemporáneo sobre la libertad de expresión, la gobernanza de internet y la concentración de poder en actores privados. Estas plataformas determinan qué voces son amplificadas, restringidas o silenciadas, lo que reconfigura la relación entre libertad de expresión, regulación y justicia. Este proceso ha

dado paso a la privatización de funciones tradicionalmente públicas, entre ellas la resolución de disputas entre las propias plataformas y los usuarios, así como la garantía de derechos fundamentales.

El Facebook Oversight Board (FOB) constituye un caso paradigmático de esta transformación. Presentado como un órgano independiente de apelación y deliberación sobre la moderación de contenidos en las plataformas de Meta, el FOB surge en un contexto de crisis de legitimidad marcado por el uso indebido de datos personales, la manipulación informativa con fines políticos y la creciente presión internacional por una mayor transparencia en sus políticas algorítmicas. Su creación representa un intento de trasladar al ámbito corporativo principios asociados al derecho internacional de los derechos humanos, al tiempo que reconfigura las fronteras entre lo público y lo privado en la administración de justicia digital.

El análisis de su funcionamiento revela, sin embargo, una tensión estructural: aunque el FOB puede emitir decisiones vinculantes en casos particulares, sus recomendaciones generales carecen de obligatoriedad. Esto limita su capacidad para incidir en la política algorítmica de Meta y refuerza la asimetría entre la empresa y los usuarios. La independencia que se le atribuye resulta, por tanto, relativa, y su papel se inscribe dentro de un modelo de autorregulación corporativa que preserva la discrecionalidad empre-

sarial bajo el lenguaje de la justicia y los derechos humanos.

El presente artículo examina el funcionamiento del FOB como dispositivo de gobernanza privada en el ecosistema digital, y evalúa sus alcances y límites en torno a la política algorítmica de la plataforma. A partir del estudio de una selección de casos emblemáticos, se analizan las tensiones entre las decisiones de la Junta, la política algorítmica de Meta y el alcance de las recomendaciones emitidas. El artículo sostiene que el FOB, pese a representar un avance en materia de rendición de cuentas, consolida un modelo de justicia privatizada que traslada la resolución de controversias sobre derechos fundamentales desde las instituciones estatales hacia estructuras corporativas de alcance global con una aplicación difusa del debido proceso.

Gobernanza de plataformas y algoritmos: gramáticas de poder en la moderación

En el ecosistema digital contemporáneo, los procesos de autorregulación y privatización de los organismos de impartición de justicia han profundizado el desplazamiento del Estado frente al poder de las grandes corporaciones tecnológicas en ámbitos clave para el desarrollo social. En este ecosistema, las plataformas sociodigitales ejercen una jurisdicción privada asumiendo funciones propias de órganos judiciales: fijan las reglas, conducen los

procedimientos y resuelven controversias dentro de la propia órbita normativa de la plataforma. A diferencia de ciertas cortes de arbitraje internacional, en las que las partes suelen tener un peso relativamente equiparable —empresas o Estados—, aquí convergen dos rasgos decisivos: una asimetría estructural de poder entre compañías y usuarios, y la incidencia directa sobre derechos fundamentales como la libertad de expresión, el acceso a la información y el desarrollo de una vida libre de violencia en el entorno digital.

Este predominio se amplifica por su integración vertical de la cadena digital —infraestructura, servicios, *software*, y algoritmos de moderación y recomendación—, que les permite orientar, restringir y condicionar la conducta en entornos que parecen foros públicos de deliberación, pero operan bajo lógicas corporativas. La configuración de estos *feudos digitales* (Celeste, 2022) marca un punto de inflexión para las políticas regulatorias, ya que las empresas han desplazado el monopolio estatal de la impartición de justicia, consolidando un orden privado-normativo en el que los dispositivos de jerarquización, la conceptualización de contenido «límitrofe» y las listas de restricciones operan como normas ejecutables sin mediación legislativa. Esta privatización se da en los marcos reglamentarios, pero también en su cumplimiento, lo que deriva en que millones de controversias sobre visibilidad y discurso se resuelven mediante procedimientos in-

ternos que raras veces alcanzan el respeto al debido proceso. Suzor (2019) denomina esto como *legalidad secreta*, una serie de normas, políticas y procedimientos internos ocultos que utilizan las grandes plataformas privadas de contenido en línea —como Facebook, Google y Twitter— para gobernar el discurso y la conducta de sus usuarios, ejerciendo un poder inmenso con una falta crítica de transparencia y rendición de cuentas.

La resistencia de las plataformas socio-digitales a una tutela regulatoria mínima sobre la moderación de contenidos se apoya en un encuadre jurídico-discursivo deliberadamente ambivalente: por un lado, se autodefinen como empresas tecnológicas o intermediarios neutrales que «solo conectan» usuarios y, por tanto, desplazan la responsabilidad de lo publicado hacia quienes postean (Gillespie, 2018; Napoli y Caplan, 2017); por otro, erigen un corpus autorregulatorio propio bajo retóricas de derechos humanos y buenas prácticas, pero sin obligaciones vinculantes ni rendición de cuentas ante autoridades externas. Esta autoidentificación como «plataformas» refuerza la idea de que no realizan actividad editorial humana ni producen contenido, aunque en la práctica sí intervienen de forma decisiva mediante procesos automatizados y algoritmos que seleccionan, filtran y jerarquizan lo que se ve, operando funcionalmente como medios que editorializan con una capacidad de intervención superior. En ese sentido, el encuadre tec-

nocéntrico señalado por Napoli y Caplan (2017) no es una mera descripción, sino un dispositivo político que invisibiliza las dimensiones culturales y públicas de su operación convirtiendo la supuesta neutralidad tecnológica en un mecanismo de exención de responsabilidad.

La propia dinámica de la moderación de contenidos desmiente, sin embargo, esa pretendida neutralidad. Las plataformas revisan, evalúan, aprueban o deniegan publicaciones según parámetros que reflejan intereses económicos y de gobernanza de la información. Determinan qué tipo de participación se promueve o desalienta, qué contenidos se muestran primero o con mayor frecuencia, cómo las herramientas de navegación guían la interacción y cómo los modelos de ingresos publicitarios organizan la información mediante sistemas de clasificación algorítmica que privilegian ciertos discursos sobre otros. La opacidad de estas jerarquías editoriales y la inexistencia de control externo sobre los criterios de priorización consolidan un espacio de gobernanza privada del discurso (Gillespie, 2018).

En paralelo, las plataformas han desarrollado regímenes normativos voluntarios fundamentados en principios que no están obligadas a cumplir. Su política de contenidos y moderación asume el carácter de «autorregulación moral»: una postura de colaboración ética frente a problemas de discurso dañino o des-

información, pero sin un marco jurídico que defina la extensión de dicha responsabilidad. Como advierte Nicolas Suzor (2019), aunque las plataformas mantienen reglas «aceptables» contra el odio y el abuso, estas reglas no se aplican de ninguna manera que pueda considerarse legítima. No hay una manera fácil de garantizar que las reglas se apliquen de manera constante o que se apliquen de una manera justa y libre de prejuicios (p. 8). La ausencia de mecanismos efectivos de aplicación y de control independiente genera un déficit estructural de legitimidad y justicia procedimental en la moderación digital.

En ese sentido, el carácter voluntario de estos compromisos, como los *Principios rectores sobre las empresas y los derechos humanos* (Naciones Unidas, 2011), refleja una autopercepción ambivalente: las plataformas se presentan como garantes de derechos y árbitros del espacio público digital, pero, al mismo tiempo, se reservan la discrecionalidad de cuándo y cómo aplicar los principios que declaran. Este funcionamiento demuestra que su compromiso con la ética y los derechos humanos opera dentro de los límites de interés corporativo, y que la autorregulación, en ausencia de supervisión pública, no sustituye la función reguladora del Estado, sino que la desplaza hacia un ámbito privado de justicia parcial y no vinculante.

En el terreno de la moderación de contenidos, esta privatización introduce dile-

mas específicos en la implementación de esquemas automatizados por algoritmos que vuelven aún más opacos los mecanismos de impartición de justicia. Sin un debido proceso en torno a la implementación algorítmica en la moderación y sus derivadas controversias —notificación comprensible, registro trazable de señales y umbrales, posibilidad real de apelación—, y sin observabilidad externa —acceso a datos para autoridades e investigadores—, el sistema deriva en justicia empresarial cuyo estándar de razonabilidad y proporcionalidad permanece indemostrado (Ananny y Crawford, 2018; Citron, 2008; Kroll *et al.*, 2017).

La aparente deshumanización del algoritmo forma parte del dispositivo discursivo de legitimación de la neutralidad tecnológica. Pero los algoritmos, diseñados por personas, responden a valores, principios e intereses determinados; clasifican, jerarquizan y visibilizan información de acuerdo con parámetros económicos y políticos que expresan la racionalidad corporativa de las compañías. Su implementación en la práctica moderadora ha permitido superar el binomio de contenido en línea o eliminado, para dar paso a una serie de medidas técnicas que permiten amplificar un contenido o disminuir su visibilidad mediante recomendaciones. La práctica de «reducir sin remover» propia de la política algorítmica ilustra por qué hace falta ese armazón regulatorio. La intervención por reducción de visibilidad se ha vuelto ubicua, evita costos

políticos de la remoción frontal y escala bien con aprendizaje automático, pero su opacidad erosiona previsibilidad y posibilidad de reparación si no está sujeta a reglas, registros y auditoría (Gillespie, 2018). Cuando esa trazabilidad existe y se somete a verificaciones independientes, la gobernanza puede evaluar sesgos sistemáticos, medir si los remedios corrigen y probar si los cambios reducen daños sin sacrificar pluralismo (Helberger, 2019; Napoli, 2011).

Esta consideración permite escapar del falso dilema entre autorregulación «buena» o «mala». Los órganos internos de las plataformas pueden aportar gran valor en la política de moderación de contenidos mediados por algoritmos —razones públicas, ordenación de criterios, corrección puntual—, pero no sustituyen la arquitectura constitucional-soberana que vuelve exigibles las obligaciones sobre la capa algorítmica (Klonick, 2018; Suzor, 2019). En este punto, el principio de la coregulación es fundamental: códigos de conducta y estándares técnicos bajo mandato público, con auditorías, *benchmarks* y remedios exigibles. El objetivo no es abolir los foros internos, sino reencuadrarlos en una arquitectura pública en la que la última palabra sobre la exposición algorítmica no dependa de un contrato de adhesión, sino de reglas verificables.

En este contexto, la constitución del Facebook Oversight Board (FOB) como junta independiente introduce elementos

novedosos, como justificación y publicidad de decisiones y un procedimiento más robusto que el *statu quo* de muchas plataformas. Sin embargo, no revierte el proceso de privatización de justicia: los estándares públicos bajo los cuales se rige están supeditados a los contratos de adhesión y a las normas comunitarias de Meta. La evaluación de riesgos sistémicos derivados del *ranking*, la trazabilidad de señales y umbrales de reducción, el acceso a datos para auditar exposición y la posibilidad de ordenar remedios estructurales son competencias que solo una arquitectura constitucional con respaldo soberano puede garantizar (Balkin, 2017; De Gregorio, 2022). Con este anclaje, el apartado siguiente describe las características del FOB —origen y mandato, independencia y financiación, tipología de casos, alcance sobre recomendación y reducción, y estatuto de sus decisiones y recomendaciones— para evaluar, a través de casos, qué aporta al gobierno de la visibilidad y dónde persiste el déficit constitucional-soberano.

El Facebook Oversight Board de Meta: arquitectura de legitimidad y privatización de la justicia digital

El surgimiento del Facebook Oversight Board (FOB) de Meta Platforms Inc. constituye uno de los experimentos institucionales más paradigmáticos del siglo XXI en materia de gobernanza digital. Tal como lo imaginaba Mark Zuckerberg antes de su constitución:

You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world [Podemos imaginar algún tipo de estructura, casi como una Corte Suprema, que esté formada por personas independientes que no trabajen para Facebook, quienes en última instancia toman la decisión final sobre lo que debería ser un discurso aceptable en una comunidad que refleja las normas y valores sociales de personas de todo el mundo] (Klein, 2018, párr. 22).

Su origen está estrechamente ligado al momento de crisis que atravesó Facebook entre 2016 y 2018, cuando los escándalos de Cambridge Analytica, la propagación de desinformación electoral y la utilización de la red para incitar violencia en Myanmar y Sri Lanka revelaron las dimensiones del poder editorial de la empresa y la opacidad de sus sistemas de moderación (Roberts, 2019). Ante la amenaza de intervención estatal y la pérdida de confianza pública, Meta optó por construir una estructura de autocontrol institucional que emulara las formas del derecho público sin ceder soberanía normativa. Así, el FOB se concibió como un mecanismo de autorregulación corporativa que, en apariencia, garantiza inde-

pendencia y pluralidad, pero que, en la práctica, funciona como una válvula de contención frente a la presión regulatoria.

Formalmente, el FOB se constituyó en 2018 mediante un fideicomiso irrevocable de 130 millones de dólares, administrado por una fundación sin fines de lucro, lo que le confiere independencia financiera, aunque no estructural. A nivel de su composición, el órgano refleja un esfuerzo por proyectar legitimidad global: reúne poco más de 20 miembros —académicos, juristas, periodistas y defensores de derechos humanos— de diversas regiones y disciplinas. Los miembros deliberan en paneles de cinco integrantes, con mandatos renovables, bajo el acompañamiento de una Secretaría Ejecutiva con sede en Londres. El proceso de selección enfatiza la diversidad geográfica y cultural, pero carece de mecanismos participativos o de escrutinio público, lo que refuerza su naturaleza cerrada y tecnocrática (Douek, 2021; Klonick, 2020). Su autoridad descansa, por tanto, en la experticia y el prestigio de sus integrantes, más que en la representación democrática.

El FOB no es un tribunal ni se origina en un poder público, sino en una empresa privada que ejerce control sobre la infraestructura global de la comunicación. Se trata de un esfuerzo corporativo por trasladar los principios de la justicia constitucional al entorno privado de las plataformas digitales. En este sentido, el FOB encarna la tensión entre la búsqueda

de legitimidad y el mantenimiento del control empresarial sobre los procesos normativos que rigen la circulación de la información dentro de una estructura autorregulatoria. Su creación es un intento de trasladar al espacio privado de las plataformas los mecanismos judiciales propios del Estado: la deliberación, el precedente, la proporcionalidad, la fundamentación jurídica y, en última instancia, la impartición de justicia.

El funcionamiento operativo del FOB reproduce la estructura de un tribunal constitucional: los usuarios presentan apelaciones por la eliminación de contenido o suspensión de cuentas y, si la respuesta no los satisface, pueden recurrir al FOB, que seleccionará aquellos casos con mayor impacto potencial o valor de precedente. Los paneles elaboran decisiones argumentadas que deben ser aprobadas por mayoría y que incluyen consideraciones de derechos humanos, referencias a tratados internacionales y razonamientos de proporcionalidad. Las resoluciones son vinculantes en el caso concreto, pero las recomendaciones sobre las normas comunitarias y las políticas algorítmicas aplicadas que acompañan cada decisión son meramente orientativas. La empresa, en última instancia, decide qué recomendaciones adopta, pospone o ignora.

La relación entre Meta y el Oversight Board revela la contradicción inherente a todo intento de institucionalizar la justicia dentro de una empresa privada.

Mientras la compañía se presenta como garante del libre discurso global, continúa siendo juez, parte y legislador de sus propias normas. Gillespie (2018) recuerda que las verdaderas decisiones editoriales de las plataformas ocurren en el nivel de los algoritmos de visibilidad, no en la superficie de la moderación. En consecuencia, aunque el FOB revisa decisiones sobre publicaciones específicas, no puede intervenir en los sistemas de clasificación algorítmica que determinan qué contenidos circulan o se ocultan, ni en los modelos de recomendación que definen la experiencia de los usuarios. Su alcance se limita al plano visible de la moderación, mientras la infraestructura técnica del poder corporativo permanece fuera de su jurisdicción.

La retórica jurídica que el Board adopta —proporcionalidad, debido proceso, derechos humanos— no debe confundirse con un marco jurídico propiamente dicho. Cohen (2019) advierte que las plataformas tienden a trasladar la semántica de los derechos hacia un marco contractual, en el que los estándares públicos se reinterpretan como políticas privadas. En ese sentido, el FOB no introduce una jurisdicción nueva, sino que busca convencer a la opinión pública y a los reguladores de que la empresa es capaz de autocontrolarse mediante procedimientos que imitan los del Estado de derecho.

El problema de fondo no es que el Board carezca de independencia total, sino que

su existencia refuerza un modelo en el cual la justicia se privatiza. Roberts (2019) sostiene que la moderación de contenido no es una forma menor de regulación, sino el nuevo rostro del poder normativo. Meta, al crear su propio tribunal, consolida la tendencia hacia la judicialización interna de la empresa: una justicia sin Estado, basada en la reputación y en el cumplimiento voluntario. Las recomendaciones del Board, aun cuando apelan a principios universales, se aplican dentro de los límites contractuales que impone la corporación. La transparencia se convierte, así, en una estrategia de legitimación, más que en un mecanismo efectivo de control.

En esa justicia privatizada, los derechos de los usuarios se interpretan como cláusulas de servicio, no como garantías constitucionales, y las sanciones adoptan la forma de gestión reputacional, más que de obligación jurídica. Aun con estas limitaciones, el FOB se ha convertido en un referente ineludible en los debates sobre gobernanza digital. Sin embargo, ese potencial regulatorio enfrenta la paradoja de su origen: un órgano privado difícilmente puede constituirse en garante universal de los derechos que el propio poder corporativo pone en tensión. La autoridad del Board proviene del mismo ecosistema económico que lo limita, y su legitimidad discursiva depende de un equilibrio precario entre crítica interna y subordinación estructural. Su valor radica menos en las

decisiones que emite que en el espejo que ofrece sobre el presente: un orden normativo global en el que las corporaciones gestionan derechos, las leyes se escriben en código y la transparencia se vuelve cada vez más opaca. Como sintetiza Suzor (2019), las plataformas digitales no son espacios fuera de la ley, sino espacios donde nuevas formas de ley privada han reemplazado silenciosamente a la pública. Comprender el FOB es entender la mutación contemporánea de la justicia, que transita de la esfera pública a la corporativa, del ciudadano al usuario y de la ley al algoritmo.

El Facebook Oversight Board y la política algorítmica de Meta

La arquitectura algorítmica de las plataformas sociodigitales constituye el eje de su régimen de gobernanza de contenidos. Su diseño optimiza métricas de compromiso, evaluando tiempo de permanencia, clics, reacciones e interacciones, y lo hace dentro del marco formal de las normas comunitarias, aunque guiado por una racionalidad económica que se traduce en intensificar la circulación, elevar la retención de audiencias y, en consecuencia, ampliar la extracción de datos y los ingresos publicitarios. Así, la decisión algorítmica no es solo una técnica de clasificación, sino una política de contenido que define qué se ve, cómo se valora y con qué velocidad se propaga, alineando la curaduría informacional con la maximización de márgenes de ganancia. Como afirma

Pickup (2021), los algoritmos encajan en esta definición de política de contenido, porque es natural referirse a un algoritmo como un procedimiento que rige el contenido de la plataforma (p. 15).

El problema se vuelve crítico cuando la lógica de amplificación algorítmica empuja a la superficie discursos de odio, xenofobia, violencia o desinformación. En ese escenario, la pregunta no es únicamente si la plataforma está dispuesta a recalibrar sus modelos para atenuar esos flujos, sino también si acepta someter su política algorítmica a un escrutinio público verificable, abrirla a la deliberación con instancias externas independientes y ejecutar las modificaciones que de ahí se deriven. Los límites institucionales son claros: el Oversight Board (FOB) no puede avanzar en propuestas vinculantes para rediseñar o imponer cambios estructurales de la plataforma; su mandato, tal como se mencionó, se circunscribe a pronunciamientos y recomendaciones cuya implementación depende, en última instancia, de la propia compañía.

De ahí que la incidencia del FOB sobre la política algorítmica de Meta sea, en términos institucionales, estrecha. Sus resoluciones se anclan en expedientes singulares y se traducen, a lo sumo, en órdenes de restituir o retirar piezas específicas de contenido. Por fuera de ese perímetro casuístico, el órgano únicamente puede formular recomendaciones generales sin fuerza obligatoria, de modo que el dise-

ño, la definición y el control de la política algorítmica permanecen como prerrogativas exclusivas de la compañía.

Ahora bien, el modo en que el FOB puede incidir sobre la arquitectura algorítmica de Meta no es mediante órdenes de rediseño, sino a través de opiniones consultivas, recomendaciones (Oversight Board, 2025) y orientación política (Oversight Board, 2024) vinculadas con la política de contenidos. Estas indicaciones carecen de fuerza vinculante y solo pueden emitirse en conexión con casos sobre los que la Junta haya resuelto; de ahí que Meta conserve discrecionalidad tanto para discutir si los algoritmos quedan comprendidos en el mandato del FOB como para implementar —o no— los ajustes sugeridos.

Con todo, y como se ha señalado, el estatuto mantiene estas opiniones consultivas, orientaciones y recomendaciones en el plano no vinculante, dejando en manos de la compañía la última palabra sobre su eventual implementación (Oversight Board, 2023a, p. 8 y p. 10). A partir de este marco normativo y de sus tensiones, conviene observar cómo se traduce en la práctica: el análisis de casos en los que el FOB ha emitido recomendaciones que tocan, de modo directo o indirecto, los procesos automatizados de moderación permite entender, con mayor precisión, hasta dónde llegan hoy las fronteras efectivas de la política algorítmica de la plataforma.

Metodología

Esta investigación adopta un enfoque cualitativo de análisis documental con lógica comparativa de casos. El diseño metodológico parte de la siguiente premisa: ante la opacidad estructural de los sistemas algorítmicos propietarios, el objeto empíricamente accesible no es el «algoritmo en sí», sino el régimen público de justificación y de respuesta institucional que Meta produce cuando el Oversight Board formula recomendaciones. En consecuencia, el estudio no pretende reconstruir el funcionamiento interno de los modelos algorítmicos, sino evaluar —en términos observables— el tipo de incidencia que el Oversight Board logra —o no logra— sobre la gobernanza de la moderación automatizada, a partir de las resoluciones oficiales que quedan registradas en el portal de transparencia corporativa.

Para construir el corpus, se realizó un mapeo sistemático de la totalidad de resoluciones y opiniones consultivas disponibles en el portal público del Oversight Board (s. f.) con corte al 31 de diciembre de 2024. En ese periodo, el FOB había publicado 150 decisiones sobre casos individuales y cuatro opiniones consultivas —Policy Advisory Opinions (PAO)—, con un total de 317 recomendaciones a Meta. Del total de resoluciones y PAO, se identificaron ocho casos en los que el expediente del Board o la respuesta de Meta hacen referencia explícita, como objeto central de análisis, a mecanismos de moderación

automatizada: sistemas de coincidencia —*matching banks* o bases de datos de huellas digitales de contenido—, algoritmos de clasificación o *ranking*, listas cerradas de gestión diferenciada —como SSR/GSR en *cross-check*—, modelos de reconocimiento de patrones textuales o visuales, o cualquier combinación de estos. De ese universo de ocho casos o PAO, la selección para el análisis comparado se realizó mediante muestreo intencional orientado por criterios sustantivos (Patton, 2002). Se incluyeron únicamente casos que satisfacían simultáneamente los cuatro criterios siguientes.

El primero es la centralidad algorítmica: el mecanismo automatizado debía ocupar un lugar sustantivo en el diagnóstico del Board, no ser un elemento periférico o meramente mencionado. El segundo es la observabilidad de la respuesta: el par recomendación-respuesta de Meta debía estar completamente publicado en el Transparency Center, con estatus de implementación actualizado, de modo que fuera posible clasificar el grado de cumplimiento reportado. El tercero es la variación de mecanismo: para que el análisis comparado aportara valor explicativo, cada caso debía exponer una arquitectura algorítmica distinta —banco de contenidos, *ranker*, listas de elegibilidad, clasificador semántico o de imagen—, evitando la redundancia analítica entre casos con lógicas técnicas similares. El cuarto es la heterogeneidad temática: se buscó que los casos cubrieran distintos tipos de con-

tenido y contexto —salud pública, crítica política, gobernanza interna del sistema, discurso sobre figuras proscritas—, de modo que los patrones identificados no fueran atribuibles a una sola categoría temática. Sobre la base de estos criterios, quedaron excluidos cuatro casos del universo acotado por redundancia analítica.

El corpus resultante comprende cuatro casos que, considerados conjuntamente, maximizan la variación de mecanismos de moderación automatizada documentados en el registro público del FOB: la opinión consultiva sobre desinformación de COVID-19, de abril de 2023 (Oversight Board, 2023b), que expone sistemas de detección por coincidencia semántica y etiquetado automático; la opinión consultiva sobre el programa *cross-check*, de diciembre de 2022 (Oversight Board, 2022b), que analiza el *ranker* GSR y las listas SSR de revisión diferenciada; el caso de la caricatura sobre la Policía colombiana, de septiembre de 2022 (Oversight Board, 2022a), que documenta el funcionamiento de los Media Matching Service *banks*; y el caso sobre Abdullah Öcalan (Oversight Board, 2021), que examina la automatización de la política sobre «dangerous individuals and organizations» [personas y organizaciones peligrosas] mediante reconocimiento facial y coincidencia de patrones textuales. El objetivo comparativo es reunir casos que expongan arquitecturas distintas de gobierno algorítmico y permitan identificar patrones transversales

de aceptación simbólica, cumplimiento procedimental y límites de penetración en el núcleo sociotécnico.

La codificación del corpus se realiza mediante análisis de contenido cualitativo orientado por reglas explícitas de decisión. Cada par recomendación-respuesta de cada caso se clasifica según (1) el tipo de problema que el Board atribuye a la automatización —error de inscripción, descontextualización, desigualdad procedimental, excepcionalidad, entre otros—, (2) el tipo de remedio propuesto —transparencia, debido proceso, evaluación de impacto, acceso a datos, cambios de política o de implementación— y (3) el grado de cumplimiento reportado por Meta.

Limitaciones metodológicas

Este enfoque asume, como condición constitutiva del objeto, la dependencia de evidencia publicada por el regulado: todo el corpus proviene de lo que Meta decide registrar y mostrar en <https://transparency.meta.com>. En consecuencia, el análisis no puede verificar técnicamente la efectividad de los cambios ni inferir con certeza modificaciones en parámetros, entrenamiento de modelos o arquitecturas internas. Del mismo modo, la plataforma puede favorecer una narrativa de cumplimiento selectivo: la empresa conserva discrecionalidad para definir el nivel de detalle, la temporalidad de las actualizaciones y el tipo de evidencia que acompaña cada estatus. Estas limitaciones no

anulan el estudio; delimitan con precisión su alcance. Lo que se ofrece es una evaluación rigurosa del régimen público de respuesta y justificación corporativa frente a recomendaciones del Oversight Board y, a partir de ello, una caracterización de los límites estructurales de este modelo de «justicia privatizada» para incidir en la gobernanza de la moderación automatizada bajo condiciones de asimetría informativa.

Resultados

Opinión consultiva sobre desinformación de COVID-19

La opinión consultiva sobre desinformación de COVID-19, publicada el 20 de abril de 2023 por el Oversight Board (2023b), representa uno de los ejercicios más reveladores sobre la interacción entre sistemas algorítmicos de moderación y principios de debido proceso digital. En este documento, Meta solicitó al Board asesoría sobre el futuro de su política de eliminación de contenidos médicos falsos tras tres años de emergencia sanitaria global. Durante ese tiempo, la empresa había aplicado un modelo mixto de automatización en detección y remoción de contenidos, basado en coincidencias semánticas, aprendizaje automático y reglas predefinidas. Millones de publicaciones fueron retiradas o degradadas mediante procedimientos en los que la intervención humana era residual y la justificación específica al usuario, prácticamente inexistente.

El FOB analizó el funcionamiento de estos sistemas y concluyó que la política de moderación aplicada durante la pandemia se encontraba en una zona gris respecto al debido proceso. Aunque respondía a una necesidad legítima de salud pública, la opacidad de las normas aplicadas y la dependencia de automatismos impedían que los usuarios comprendieran con claridad qué regla habían violado, cómo se evaluó su publicación y qué vía real tenían para apelar. Desde una perspectiva de derechos humanos, este déficit procedimental afectaba la previsibilidad de la norma, la oportunidad de defensa y la proporcionalidad de la sanción.

Por esa razón, el FOB emitió más de 18 recomendaciones dirigidas a corregir estos vacíos estructurales. Entre ellas destacó la solicitud de realizar una evaluación de impacto en derechos humanos (HRIA) sobre los algoritmos de recomendación y *ranking* para determinar si amplificaban o mitigaban la desinformación sanitaria; la exigencia de publicar de forma transparente las aproximadamente 80 afirmaciones médicas sujetas a retirada y sus justificaciones específicas; la creación de mecanismos de apelación con mayor información para el usuario; y la habilitación de acceso a datos para investigación independiente sobre la eficacia de las etiquetas y verificaciones.

Sin embargo, la respuesta de Meta al documento del Board fue ambivalente (Meta, 2023c). En su portal del Trans-

parency Center, la empresa aceptó parcialmente varias recomendaciones, pero desestimó otras bajo el argumento de que la Organización Mundial de la Salud había declarado el fin de la emergencia internacional por la COVID-19 el 5 de mayo de 2023. Este cambio de contexto le sirvió como fundamento para no continuar con las revisiones que el Board había solicitado. Las recomendaciones incluidas en el punto 1 de la opinión consultiva relativas a la revisión periódica y consulta pública (Oversight Board, 2023b) fueron marcadas como «no further action» [ninguna acción adicional]. Meta consideró que ya no era pertinente seguir publicando el listado completo de afirmaciones prohibidas ni realizar procesos de consulta abierta para revisarlas, aunque el FOB había pedido precisamente que esa práctica continuara como ejemplo de buena gobernanza algorítmica más allá de la contingencia.

Una segunda área de ambivalencia se observa en la recomendación 10, relativa a la evaluación de impacto en derechos humanos sobre los sistemas de recomendación y distribución de contenidos. Meta clasificó esta acción como en fase de *evaluación de la viabilidad*, sin calendario ni parámetros definidos. El lenguaje utilizado en su respuesta —«We are still in the early phases of assessing the feasibility of this complex recommendation, and are currently working with a number of internal teams including Human Rights, Policy, Research teams, and others» [Aún

nos encontramos en las primeras fases de evaluación de la viabilidad de esta compleja recomendación y actualmente colaboramos con diversos equipos internos, incluidos los de Derechos Humanos, Políticas, Investigación y otros] (Meta, 2023c)— deja la decisión sujeta a criterios internos no verificables. Esta ambigüedad contrasta con el espíritu de la recomendación del FOB, que exigía una auditoría específica sobre la arquitectura algorítmica de amplificación de contenido médico falso.

La recomendación 16 también presenta un caso claro de ambivalencia. En ella, el Board pidió publicar los resultados del estudio interno sobre la eficacia de las etiquetas neutras y de verificación, para permitir la evaluación independiente de su impacto. Meta afirmó haber compartido «raw findings» [hallazgos en bruto] con el propio Board, pero se negó a divulgar los resultados completos por motivos de confidencialidad y sensibilidad de datos, marcando la recomendación igualmente como «no further action» [ninguna acción adicional]. De este modo, reconoció la relevancia del conocimiento, pero mantuvo su carácter reservado, limitando la posibilidad de escrutinio público.

En suma, el caso COVID-19 constituye una muestra precisa de cómo la relación entre el FOB y Meta oscila entre la colaboración y la ambigüedad. Las recomendaciones más directamente relacionadas con la transparencia algorítmica, la evaluación

externa y la trazabilidad de decisiones, es decir, aquellas que tocan el corazón del debido proceso digital, fueron precisamente las que quedaron sin implementación plena. Esta ambivalencia evidencia que la arquitectura técnica de la moderación sigue siendo un ámbito de soberanía corporativa que escapa al control del órgano que, en teoría, fue creado para supervisarla.

Opinión consultiva sobre el sistema *cross-check* de Meta

La opinión consultiva sobre el programa *cross-check* emitida por el Oversight Board en 2022 (Oversight Board, 2022b) es otro ejemplo fundamental para comprender el funcionamiento interno de la moderación algorítmica en Meta y las tensiones que emergen cuando una empresa privada asume funciones cuasi-jurisdiccionales sobre el discurso público digital. El programa *cross-check*, establecido originalmente por Facebook, tiene como propósito corregir errores de sobreaplicación de las normas —evitar que publicaciones legítimas de alto impacto sean removidas de manera automática—, pero, en la práctica, ha creado un sistema paralelo de moderación privilegiada cuya arquitectura se sostiene sobre dos pilares técnicos: las listas cerradas de revisión secundaria (SSR) y el algoritmo de priorización o ranker (GSR).

El FOB describió que las listas SSR agrupan cuentas y organizaciones considera-

das de «alto valor» o «interés público», entre ellas medios de comunicación, funcionarios gubernamentales, celebridades y socios comerciales. Cuando el contenido de estas cuentas es marcado por los sistemas automatizados como potencialmente infractor, no se ejecuta la medida estándar de forma inmediata, sino que se detiene el flujo para una revisión reforzada por parte de un equipo especializado. Paralelamente, el algoritmo *cross-check ranker* evalúa todas las publicaciones detectadas por los clasificadores automáticos y calcula, mediante un conjunto de variables ponderadas, la probabilidad de error, la gravedad de la posible infracción, el alcance previsto del contenido y la sensibilidad temática o contextual. Con base en esa puntuación, decide si la publicación debe ingresar a la revisión secundaria y con qué prioridad.

En teoría, esta arquitectura busca aumentar la precisión de la moderación y reducir los falsos positivos. Sin embargo, el FOB observó que el sistema genera diferencias de trato entre usuarios y que la lógica de priorización reproduce sesgos estructurales: los criterios de inclusión en las listas privilegian a actores con poder político o económico, mientras que la evaluación automatizada del *ranker* carece de mecanismos de auditoría externa. El resultado, según el Board, es un esquema de moderación que preserva la asimetría informativa entre la empresa y los usuarios, y que carece de los componentes básicos del debido proceso,

como notificación efectiva, oportunidad de defensa y transparencia sobre la regla aplicada.

Ante este diagnóstico, el FOB formuló 32 recomendaciones dirigidas a Meta, entre ellas varias de carácter estructural que implicaban modificaciones directas en su política algorítmica. Solicitó, en primer lugar, la creación de un mecanismo público de postulación para que cualquier usuario pudiera solicitar la protección de *cross-check* cuando existieran riesgos de sobremoderación vinculados con derechos humanos o contextos de vulnerabilidad, no solo cuando la empresa considerara valioso el perfil. De esta forma, se buscaba reducir la dependencia del programa respecto a listas cerradas y avanzar hacia un esquema más equitativo.

En segundo lugar, el Board recomendó notificar a todas las cuentas incluidas en las listas y ofrecer una opción de exclusión voluntaria —*opt-out*—, de modo que los usuarios pudieran conocer y controlar el tipo de revisión que reciben. A ello sumó la sugerencia de marcar públicamente las categorías de entidades que gozan de revisión reforzada —medios de comunicación o actores estatales— para que la ciudadanía comprenda cuándo un contenido se encuentra sujeto a un proceso diferenciado.

En el plano algorítmico, el FOB exigió también que Meta desarrollara y publicara métricas de desempeño del programa,

en particular el número de vistas adicionales que obtienen las publicaciones infractoras mientras esperan revisión. Esa métrica es clave para determinar el efecto adverso del retardo algorítmico: cuánto daño potencial se genera al mantener visible un contenido dañino bajo la justificación de revisión especial.

Finalmente, el FOB planteó la necesidad de separar explícitamente las vías de revisión que responden a criterios de derechos humanos de aquellas motivadas por intereses comerciales. El algoritmo de priorización debía ponderar con mayor peso los contenidos con relevancia pública o impacto social, en lugar de priorizar el valor reputacional o económico del emisor. Esta separación, además de una medida ética, pretendía corregir una distorsión técnica: el *ranker* estaba calibrado para reducir el riesgo de error según la importancia del actor, no según el valor social del mensaje, lo que reforzaba una jerarquía de visibilidad dentro del sistema automatizado.

La respuesta de Meta (2024) a estas recomendaciones fue, en gran parte, negativa o ambivalente. En su comunicación oficial, la empresa aceptó parcialmente mejorar los tiempos de revisión y documentar el funcionamiento del algoritmo, pero rechazó implementar los cambios estructurales propuestos. En concreto, Meta no estableció un proceso público de postulación a *cross-check*, no notifica a las cuentas incluidas en las listas, no

marca públicamente las categorías de entidades privilegiadas y no publica métricas verificables sobre el efecto de los retrasos. Todas estas recomendaciones fueron clasificadas como «no further action» [ninguna acción adicional] en el tablero de seguimiento del Transparency Center. La negativa de Meta a marcar públicamente los contenidos o actores sujetos a revisión reforzada impide a la ciudadanía identificar cómo y cuándo intervienen las máquinas; su decisión de no publicar métricas de vistas adicionales mantiene opaca la magnitud del daño derivado de los retrasos; y su rechazo a abrir procesos de inclusión y exclusión transparentes consolida un régimen de elegibilidad cerrado que contradice el principio de igualdad procedimental.

En síntesis, la moderación *cross-check* combina decisiones humanas y automatizadas en una arquitectura de jerarquías invisibles. El Oversight Board propuso convertir ese sistema en un mecanismo transparente, auditable y centrado en derechos humanos; Meta, en cambio, se limitó a optimizar su eficiencia interna y a conservar la discrecionalidad sobre los flujos algorítmicos que determinan qué contenidos merecen una segunda revisión y cuáles no. El resultado es un modelo de gobernanza asimétrica, en el que el debido proceso se fragmenta: existen dos niveles de justicia algorítmica, uno reforzado para los actores privilegiados y otro automatizado para el resto de los usuarios.

Caricatura sobre la Policía colombiana

El caso de la caricatura sobre la Policía colombiana (Oversight Board, 2022a) permite observar con nitidez cómo una decisión errónea en la fase humana puede amplificarse de forma masiva cuando entra en juego la automatización. Un usuario en Colombia publicó, como comentario, una imagen satírica que parodiaba el escudo de la Policía Nacional para denunciar violencia policial en contexto de protesta. Dieciséis meses después, Meta removió el contenido porque coincidía con una imagen de su banco de contenidos en sus Media Matching Service (MMS) *banks* —repositorios que automatizan acciones sobre imágenes previamente catalogadas— bajo la norma de «dangerous individuals and organizations» [personas y organizaciones peligrosas].

La remoción no se apoyó en un nuevo juicio sustantivo sobre la publicación, sino en el arrastre mecánico de una decisión previa: la imagen ya estaba etiquetada en el banco como si representara a un actor peligroso. La intervención del FOB revirtió la eliminación y calificó de problemática la arquitectura de los MMS *banks*: un solo error desencadena remociones automáticas en cadena de copias idénticas, incluso cuando el contenido es lícito y de interés público. En términos de debido proceso, esto erosiona la previsibilidad de la regla aplicable y la oportunidad real de defensa del usuario, que desconoce el

porqué de la medida y enfrenta un circuito en el que la máquina replica un error humano a gran escala.

El expediente muestra tres elementos clave sobre el funcionamiento técnico de estos bancos. Primero, el uso de bancos de información convierte la imagen en un contenido asociado a una política específica —por ejemplo, «dangerous individuals and organizations» [personas y organizaciones peligrosas], «hate speech» [discurso de odio] o «child sexual exploitation» [explotación sexual infantil]—; a partir de ahí, el sistema puede borrar copias al momento de subirlas, escanear y retirar material ya publicado, añadir pantallas de advertencia, o servir de señal a los revisores humanos. Segundo, la entrada de contenidos a los bancos depende de revisión humana a escala, con salvaguardas como la Dynamic Multi-Review —dos revisores deben coincidir—, pero, como admite la propia Meta, esas barreras no impidieron que la caricatura fuera colocada en un banco de datos como si violara la política de organizaciones criminales. Tercero, aunque existen alertas de anomalías —por ejemplo, muchas apelaciones exitosas—, no se activó a tiempo una depuración del banco: 215 usuarios apelaron y el 98 % ganó, pero la imagen siguió en el banco hasta que el Board tomó el caso; solo entonces Meta restauró ese y otros contenidos idénticos que habían sido removidos por la misma causa. Este retraso procesal —entre la señal de que el banco se estaba equivocando y

la limpieza efectiva del banco— es, para el FOB, el corazón del déficit de debido proceso automatizado: si un mecanismo de corrección no «saca» con rapidez el contenido lícito, el sistema continúa multiplicando la remoción errónea y obstaculizando apelaciones significativas.

Además de revisar la consistencia con las Normas Comunitarias —la propia Meta admitió que el dibujo no violaba ni el estándar comunitario de *organizaciones peligrosas* ni el de *violencia e incitación*, y precisó que la regla de «violent and graphic content» [contenido violento y gráfico] no se aplica a caricaturas—, el Board encuadró el análisis en estándares de derechos humanos: la sátira política y la crítica a violaciones de derechos por parte de fuerzas estatales gozan de máxima protección bajo el artículo 19 del Pacto Internacional de Derechos Civiles y Políticos; cualquier restricción, incluso en un marco privado, debe ser necesaria y proporcional al objetivo de protección.

Al colocar en bancos de datos y retirar automáticamente múltiples copias de una crítica a agentes estatales, sin que mediara una valoración sustantiva ni vías efectivas de corrección, el sistema incurrió en una medida desproporcionada, incompatible con las responsabilidades de Meta en materia de libertad de expresión. El Board también señaló un ángulo de transparencia: Meta impidió divulgar el número total de remociones ligadas al uso de ese banco de datos, lo que dificulta

medir el alcance colectivo del error y fijar metas de reducción.

Las recomendaciones del FOB fueron precisas y orientadas a mecanismos: crear un mecanismo de reevaluación para «that content with high rates of appeal and high rates of successful appeal is reassessed for possible removal from its Media Matching Service banks» [que el contenido con altos índices de atractivo y altas tasas de éxito en las solicitudes sea reevaluado para su posible eliminación de los bancos de su Servicio de Coincidencia de Medios] (Oversight Board, 2022a); fijar estándares de tiempo entre la detección de un posible error y la eliminación del contenido del banco cuando se compruebe que no infringe; y publicar tasas de error por política —desagregadas por estándar— en los informes de transparencia, con información sobre cómo entra contenido a los bancos y qué medidas toma la empresa para reducir fallos. Estas pautas buscan convertir la apelación exitosa en una señal fuerte que limpie el banco y que, por tanto, detenga la reproducción del error por la capa automatizada.

En esta línea, la plataforma (Meta, 2023b) rechazó —o diluyó sin plazos— lo más sustantivo: la publicación de métricas públicas de error de los bancos de contenidos, delimitar estándares de tiempo para «limpiar» los errores de estos bancos y una mayor transparencia operativa sobre cómo se crean, auditan y depuran estos bancos, rechazo que incluye la notificación significa-

tiva al usuario cuando una acción se debe a una coincidencia de banco.

Desde la óptica del debido proceso digital, el caso deja tres aprendizajes. Primero, la legalidad operativa: la regla no solo debe ser clara en abstracto; también debe existir una cadena de trazabilidad entre la decisión de banqueo, la acción automática y la rectificación oportuna cuando hay evidencia de error. Segundo, la notificación significativa: si el usuario recibe mensajes genéricos, sin explicar que la remoción provino de una coincidencia con el banco de contenidos y no de una nueva evaluación, su capacidad de defensa se vacía; del mismo modo, si el sistema sigue retirando copias mientras las apelaciones prosperan, la oportunidad de ser oído queda en papel. Tercero, la proporcionalidad técnica: sistemas que multiplican una sola decisión a toda la plataforma exigen métricas específicas de error y corrección; sin estas métricas, el riesgo de sobrecumplimiento automatizado contra discurso protegido, como la denuncia de abuso policial, es estructural.

El aislamiento de Öcalan

El caso del aislamiento de Öcalan (Oversight Board, 2021) enfrentó al FOB con uno de los dilemas más profundos de la gobernanza algorítmica: cómo equilibrar la seguridad pública y la libertad de expresión política en entornos automatizados de moderación. El contenido original fue una publicación en Facebook que repro-

ducía la fotografía del líder kurdo Abdullah Öcalan, fundador del Partido de los Trabajadores del Kurdistan (PKK), organización clasificada por varios Estados, incluida Turquía, como grupo terrorista, acompañada de un texto que llamaba a su liberación y lo describía como figura clave del movimiento kurdo. Meta eliminó el contenido automáticamente bajo su política de «dangerous individuals and organizations» [personas y organizaciones peligrosas], al considerar que la publicación constituía elogio o apoyo a una entidad proscrita.

El FOB, tras examinar el caso, revocó la decisión y ordenó restaurar el contenido, sosteniendo que la política de Meta y su aplicación algorítmica eran demasiado amplias y descontextualizadas. Según el Board, la publicación no promovía violencia ni incitaba a actos terroristas, sino que expresaba una opinión política legítima sobre la detención de un líder político. La eliminación, por tanto, resultaba incompatible con los estándares internacionales de libertad de expresión, particularmente el artículo 19 del Pacto Internacional de Derechos Civiles y Políticos, que protege la discusión de temas de interés público, incluidos los relativos a movimientos separatistas o conflictos armados.

En el plano técnico, el caso reveló cómo las políticas de las consideradas *personas y organizaciones peligrosas* operan mediante sistemas de coincidencia algo-

rítmica. Meta mantiene una lista interna de individuos y entidades prohibidas, actualizada con base en designaciones estatales y fuentes propias, y los clasificadores automatizados detectan nombres, símbolos e imágenes asociados para bloquear o degradar contenido de forma preventiva. En este caso, la detección fue completamente automática: el sistema identificó el rostro de Öcalan —registrado en su base de datos de «dangerous individuals» [personas peligrosas]— y activó la retirada inmediata del post. Ninguna revisión humana intervino antes de la acción. El FOB observó que este tipo de automatización ignora contexto, intención y discurso crítico o informativo, y que, al no existir canales efectivos de apelación ni explicación significativa al usuario, la moderación se convierte en un acto de censura algorítmica opaca.

El Board señaló, además, una inconsistencia estructural: mientras que las políticas de Meta prohíben «elogiar» a personas o grupos designados, no diferencian entre incitación a la violencia y discusión política legítima. Esta confusión se agrava cuando la ejecución de las normas recae en sistemas automatizados entrenados para detectar patrones de coocurrencia —imágenes, nombres, palabras clave— sin evaluación semántica o contextual. Así, la mención o representación visual de un líder proscrito, incluso en contextos periodísticos, educativos o políticos, es tratada del mismo modo que un mensaje de reclutamiento o apología armada.

En sus recomendaciones, el FOB pidió a Meta reformar integralmente la política de «dangerous individuals and organizations» [personas y organizaciones peligrosas], incorporando excepciones explícitas para contenidos periodísticos, académicos y de derechos humanos que aludan a individuos o grupos designados sin promover su violencia. Recomendó también revisar los criterios algorítmicos de detección —en particular, los modelos de reconocimiento facial y coincidencia de texto— para garantizar que no se activen automáticamente en ausencia de un contexto incitatorio, y que se introduzcan procesos de revisión humana obligatoria en todos los casos en los que haya duda sobre la naturaleza política o informativa del discurso.

La respuesta de Meta fue ambivalente (Meta, 2023a). La empresa aceptó restaurar el contenido y reconoció la validez de proteger discursos sobre figuras políticas, pero no adoptó plenamente las recomendaciones estructurales. En su informe de cumplimiento, Meta (2021) señaló: «We are in the process of reviewing updated language to our policy rationale to ensure it accurately reflects the ways we consider human rights violations and abuse are a part of our Dangerous Organizations and Individuals policy» [Estamos en proceso de revisar la redacción actualizada de la justificación de nuestra política para garantizar que refleje con precisión la forma en que consideramos que las violaciones y abusos

de los derechos humanos forman parte de nuestra política sobre Organizaciones e Individuos Peligrosos] (p. 28). Sin embargo, rechazó modificar los sistemas de detección automatizada: «We will take no further action on this recommendation because the policy guidance in this case does not directly contribute to the performance of automated enforcement» [No tomaremos ninguna otra medida con respecto a esta recomendación, ya que la guía política en este caso no contribuye directamente al desempeño de la aplicación automatizada] (Meta, 2023a). En consecuencia, el proceso de revisión de coincidencias faciales y textuales sigue automatizado, y las excepciones contextuales quedan delegadas al juicio del revisor posterior, si el usuario apela.

En conjunto, estas recomendaciones y respuestas delimitan un mapa claro sobre cómo la moderación algorítmica de Meta opera bajo observación externa. El FOB insiste en la necesidad de transparencia, trazabilidad y corrección humana en procesos dominados por automatismos. Meta, por su parte, adopta gradualmente algunas de estas sugerencias, pero conserva el control sobre la arquitectura técnica que determina qué ve, publica o puede apelar cada usuario. En este punto, se evidencia que el debate sobre el debido proceso en la moderación digital no se li-

mita al derecho a apelar, sino que abarca la estructura misma de los sistemas algorítmicos que deciden quién tiene voz en el espacio público digital.

Discusión y conclusiones: autogobierno corporativo, límites de la política algorítmica y el alcance real del FOB

Leídos en conjunto —el caso de Abdullah Öcalan, el de la caricatura sobre la Policía colombiana, y las opiniones consultivas sobre *cross-check* y la COVID-19—, los casos exhiben un patrón estructural: la moderación de Meta descansa en dispositivos algorítmicos de escala —listas, *rankers*, bancos de coincidencia, clasificadores— que optimizan velocidad y cobertura, pero empobrecen el juicio contextual, multiplican errores cuando la política algorítmica es defectuosa y mantienen opaca la lógica decisoria ante usuarios, investigadores y el propio Oversight Board (FOB). Allí donde el Board propuso introducir previsibilidad, trazabilidad y auditoría, la respuesta corporativa fue aceptar ajustes de proceso o mensajería, sin ceder control sobre el núcleo técnico de la visibilidad y la sanción.

Desde los principios del debido proceso algorítmico², los cuatro casos convergen en tres déficits recurrentes:

²El debido proceso algorítmico es el conjunto de garantías procedimentales que deben acompañar las decisiones mediadas por sistemas automatizados —modelos de clasificación, reglas computacionales o mecanismos de moderación a escala— cuando esas decisiones afectan derechos o intereses de las personas, de modo que existan razones identificables y comunicables, criterios trazables, oportunidad real de impugnación, revisión humana significativa y vías efectivas de corrección o reparación frente a errores o sesgos.

- Inteligibilidad normativa: usuarios y terceros no disponen de reglas operativas verificables sobre qué dispara la intervención, cómo se pondera el riesgo y qué efectos secundarios —reducciones de alcance, rotulados, rebanqueo— se aplican y durante cuánto tiempo.
- Trazabilidad y corrección: las vías de reclamación no siempre modifican con rapidez la causa técnica del error —por ejemplo, eliminar el *hash* del banco, recalibrar el umbral del clasificador, mover de inmediato un caso a revisión reforzada por contexto de protesta o interés público—.
- Observabilidad externa: sin métricas públicas sobre precisión diferencial de *cross-check* o sobre el daño por retraso, sin evaluación de impacto en derechos humanos del *ranking*, y sin publicación de estudios sobre etiquetas, el sistema moderador carece de rigurosidad.

A la luz de esto, el FOB introduce lenguaje, estándares y correcciones puntuales valiosas —restaura discursos injustamente retirados, empuja hacia el etiquetado proporcional, recomienda métricas y avisos—, pero su fuerza vinculante se agota en el caso individual; en lo sistémico, sus recomendaciones son no obligatorias y dependen de la voluntad de la empresa. La combinación de mandato acotado, de-

pendencia estructural y falta de acceso técnico impide que el Board actúe como autoridad regulatoria en sentido estricto: no puede imponer auditorías del *ranking* y tampoco puede ordenar la publicación de métricas críticas ni controlar el diseño de listas y bancos.

Los déficits propios de las entidades privadas de «impartición de justicia» que atraviesan este cuadro refuerzan el diagnóstico. Carecen de legitimidad democrática de origen: sus miembros no emanan de procesos electorales ni de mandatos públicos, y su autoridad descansa en la experticia y en contratos privados, no en una jurisdicción soberana. Su independencia es formal pero no estructural: dependen del financiamiento, del mandato y del acceso que la propia empresa concede; esto crea incentivos de captura —de agenda y de información— y zonas de conflicto de interés difíciles de disipar. El derecho de defensa es limitado: no hay acceso a la evidencia técnica relevante —códigos, umbrales, listas internas—, y la igualdad entre usuario y plataforma es, en la práctica, inexistente por la asimetría de datos.

La fuerza ejecutiva de las resoluciones es acotada: pueden obligar en el caso puntual —retirar/restaurar—, pero no rediseñar sistemas ni imponer remedios estructurales, y el cumplimiento de recomendaciones generales es voluntario, ya que la empresa puede derivar o reinterpretar criterios a conveniencia. En

ese marco, los «avances» más tangibles se logran cuando la recomendación no exige abrir la caja negra. Mientras la arquitectura algorítmica central permanece cerrada y las limitaciones inherentes a la «justicia privada» impiden convertir sus diagnósticos en remedios estructurales, las recomendaciones del FOB en torno a la explicabilidad y control externo —auditorías del *feed*, métricas de retraso, marcado público de categorías privilegiadas, publicación de estudios— son totalmente deficientes.

La lección transversal es clara: sin controles públicos obligatorios sobre el diseño y las métricas de los sistemas de ampliación y sanción —y sin un andamiaje institucional que otorgue legitimidad, independencia material, publicidad y ejecutoriedad—, la autorregulación no corregirá los fallos de constitucionalidad práctica del espacio público digital: igualdad de trato, previsibilidad y posibilidad real de defensa ante máquinas que deciden a escala.

El Facebook Oversight Board representa un experimento significativo, pero insuficiente. Demuestra que es posible introducir deliberación razonada, precedentes argumentados y principios de derechos humanos en la moderación corporativa, pero también revela que estos avances no pueden sustituir la regulación pública cuando la arquitectura algorítmica que gobierna la visibilidad del discurso permanece bajo control exclusivo de la

empresa. La justicia privatizada, por sofisticada que sea su construcción institucional, carece de la legitimidad democrática, la independencia estructural y los instrumentos coercitivos necesarios para garantizar que los derechos fundamentales en el espacio digital se ejerzan con igualdad y se protejan con efectividad.

El desafío contemporáneo no consiste en elegir entre autorregulación y regulación estatal, sino en construir modelos de coregulación en los que los mecanismos internos de las plataformas —incluyendo órganos como el FOB— operen dentro de un marco normativo público que establezca estándares mínimos verificables, habilite auditorías independientes, garantice acceso a datos para investigación y otorgue a las autoridades competentes la capacidad de imponer remedios estructurales cuando los sistemas algorítmicos vulneren derechos a escala.

Solo así será posible transitar de la justicia privatizada hacia una gobernanza algorítmica genuinamente democrática, en la que el poder de decidir qué se ve y qué se silencia en el espacio público digital no dependa exclusivamente de la voluntad de las corporaciones, sino de reglas públicas, transparentes y sometidas al escrutinio ciudadano.

REFERENCIAS

- Ananny, M. y Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. <https://doi.org/10.1177/1461444816676645>
- Balkin, J. M. (2018). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davies Law Review*, 51(3), 1149-1210. <https://lawreview.law.ucdavis.edu/archives/51/3/free-speech-algorithmic-society-big-data-private-governance-and-new-school-speech>
- Celeste, E. (2022). *Digital constitutionalism: The role of internet bills of rights*. Routledge. <https://doi.org/10.4324/9781003256908>
- Citron, D. K. (2008). Technological due process. *Washington University Law Review*, 85(6), 1249-1313. <https://journals.library.wustl.edu/lawreview/article/id/6697/>
- Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press.
- De Gregorio, G. (2022). *Digital constitutionalism in Europe: Reframing rights and powers in the algorithmic society*. Cambridge University Press. <https://doi.org/10.1017/9781009071215>
- Douek, E. (2021). The limits of international law in content moderation. *UC Irvine Journal of International, Transnational, and Comparative Law*, 6(1), 37-76. <https://escholarship.org/uc/item/2857f1jq>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), 993-1012. <https://doi.org/10.1080/21670811.2019.1623700>
- Klein, E. (2018, 2 de abril). *Mark Zuckerberg on Facebook's hardest year, and what comes next*. Vox. <https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge>
- Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6), 1598-1670. <https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/>
- Klonick, K. (2020). The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal Forum*, 129, 2418-2499. <https://yalelawjournal.org/feature/the-facebook-oversight-board>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. y Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705.
- Meta. (2021, noviembre). *Meta Q2 + Q3 2021 quarterly update on the Oversight Board*. <https://about.fb.com/wp-content/uploads/2021/11/Meta-Q2-and-Q3-2021-Quarterly-Update-on-the-Oversight-Board.pdf>

- Meta. (2023a, 12 de junio). *Case regarding the support of Abdullah Öcalan, founder of the PKK*. Meta Transparency Center. <https://transparency.meta.com/oversight/oversight-board-cases/support-of-abdullah-ocalan-founder-of-the-pkk>
- Meta. (2023b, 12 de junio). *Oversight Board selects a case regarding a comment with a cartoon depicting violence by the Police in Colombia*. Meta Transparency Center. <https://transparency.meta.com/oversight/oversight-board-cases/cartoon-depicting-violence>
- Meta. (2023c, 20 de junio). *Oversight Board selects a PAO on the removal of COVID-19 misinformation*. Meta Transparency Center. <https://transparency.meta.com/oversight/oversight-board-cases/pao-on-COVID-19-misinformation/>
- Meta. (2024, 3 de octubre). *Oversight Board selects a PAO on Meta's cross-check policies*. Meta Transparency Center. <https://transparency.meta.com/oversight/oversight-board-cases/pao-cross-check-policy/>
- Naciones Unidas. (2011). *Principios rectores sobre las empresas y los derechos humanos. Puesta en práctica del marco de las Naciones Unidas para «proteger, respetar y remediar»*. https://www.ohchr.org/sites/default/files/documents/publications/guiding-principlesbusinesshr_sp.pdf
- Napoli, P. M. (2011). *Audience evolution: New technologies and the transformation of media audiences*. Columbia University Press.
- Napoli, P. y Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, 22(5). <https://doi.org/10.5210/fm.v22i5.7051>
- Oversight Board. (s. f.). *Case decisions and policy advisory opinions*. <https://www.oversightboard.com/decision/>
- Oversight Board. (2021, 8 de julio). *Öcalan's isolation*. <https://www.oversightboard.com/decision/IG-I9DP23IB/>
- Oversight Board. (2022a, 15 de septiembre). *Colombian Police cartoon*. <https://www.oversightboard.com/decision/FB-I964KKM6/>
- Oversight Board. (2022b, 6 de diciembre). *Oversight Board publishes policy advisory opinion on meta's cross-check program*. <https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/>
- Oversight Board. (2023a, febrero). *Oversight Board charter*. <https://www.oversightboard.com/wp-content/uploads/2023/11/3427086457563794.pdf>
- Oversight Board. (2023b, 20 de abril). *Removal of COVID-19 misinformation*. <https://www.oversightboard.com/decision/pao-sabu4p2s/>
- Oversight Board. (2024). *Rulebook for case review and policy guidance*. https://www.oversightboard.com/wp-content/uploads/2024/03/OB_Rulebook_March_2024-1.pdf

- Oversight Board. (2025, junio). *Oversight Board bylaws*. <https://www.oversightboard.com/wp-content/uploads/2025/07/Oversight-Board-Bylaws-June-2025.pdf>
- Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3.^a ed.). Sage Publications.
- Pickup, E. L. (2021). The Oversight Board's dormant power to review Facebook's algorithms. *Yale Journal on Regulation Bulletin*, 39(1). <https://www.yalejreg.com/bulletin/the-oversight-boards-dormant-power-to-review-facebooks-algorithms/>
- Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.
- Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.

Autor correspondiente: Dardo Emanuel Neubauer
(dardoneubauer@politicas.unam.mx)

Roles de autor: Neubauer, D.: conceptualización; metodología; investigación; escritura - borrador original; escritura, revisión y edición; supervisión y administración del proyecto

Cómo citar este artículo: Neubauer, D. E. (2026). Justicia privatizada y gobernanza algorítmica: el alcance real del Facebook Oversight Board. *Conexión*, (25), 43-72.
<https://doi.org/10.18800/conexion.202601.002>

Este es un artículo de acceso abierto distribuido bajo los términos de Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0), que permite el uso, la distribución y la reproducción sin restricciones en cualquier medio, siempre que se cite correctamente la obra original.