https://revistas.unitec.edu/tekne
Vol. 4, No. 2, 2026
ISSN: 2959-6629
OPEN ACCESS

Original

Inteligencia artificial, discurso de odio y libertad de expresión: estándares comparados de la Corte IDH y el TEDH

Artificial Intelligence, Hate Speech, and Freedom of Expression: A Comparison of the Standards of the Inter-American Court of Human Rights and the European Court of Human Rights

Ines Yadira Cubero González*[1]

[1]Facultad de Ciencias Administrativas y Sociales, Universidad Tecnológica Centroamericana, UNITEC, San Pedro Sula, Honduras

*Autor corresponsal: cuberoines@unitec.edu

Este trabajo está bajo una licencia internacional Creative Commons Attribution 4.0 BY, NC.

Recibido: 28 de marzo de 2026 / Aceptado: 11 de junio de 2026 / Publicado: 15 de septiembre de 2026.
Citar: Cubero González, I. Y. (2026). Inteligencia artificial, discurso de odio y libertad de expresión: estándares comparados de la Corte IDH y el TEDH. *Tekné: Ciencias Sociales y Humanidades, 4*(2), 1-11. https://doi.org/10.69845/tekn.v4i2.608

Resumen. Introducción. Introducción. El presente trabajo analiza la arquitectura jurídico-fundacional de la moderación algorítmica a partir de la tensión antinómica entre libertad de expresión y discurso de odio. La relevancia de este análisis radica en la necesidad de desentrañar desde una perspectiva dogmática, la disrupción categorial que provoca la inteligencia artificial generativa como un mero instrumento técnico dentro del Derecho Internacional. **Métodos**. Se realizó una revisión crítica y comparativa de la jurisprudencia reciente de la Corte Interamericana de Derechos Humanos y el Tribunal Europeo de Derechos Humanos. **Resultados.** Los hallazgos revelan una divergencia de estándares: mientras el sistema europeo consolida un modelo de *due diligence* que expande la responsabilidad penal hacia el titular del perfil y las plataformas, el sistema interamericano mantiene una resistencia estructural frente a toda forma de censura preventiva en debates de interés público. **Conclusión.** A partir de las tesis de Cotino Hueso y Llansó, se concluye que la responsabilidad por esta codificación regulatoria en el ecosistema digital recae directamente en los dueños de las redes sociales y en los perfiles particulares que violentan las normas, más que en la capacidad de la IA en sí misma. La delegación del juicio al algoritmo binario genera opacidad y una "limpieza higiénica" que sacrifica el pluralismo democrático al subordinar los derechos fundamentales a una lógica de neutralización del daño lingüístico, desprovista de hermenéutica jurídica.
Palabras Clave: Corte Interamericana de Derechos Humanos, Discurso de odio, Inteligencia artificial, Libertad de expresión, Moderación de contenidos, Tribunal Europeo de Derechos Humanos

Abstract. Introduction. This paper analyzes the legal and foundational framework of algorithmic moderation based on the antinomic tension between freedom of expression and hate speech. The relevance of this analysis lies in the need to unravel, from a dogmatic perspective, the categorical disruption caused by generative artificial intelligence as a mere technical tool within international law. **Methods.** A critical and comparative review of recent case law from the Inter-American Court of Human Rights and the European Court of Human Rights was conducted. **Results.** The findings reveal a divergence in standards: while the European system consolidates a due diligence model that extends criminal liability to account holders and platforms, the Inter-American system maintains a structural resistance to any form of preventive censorship in debates of public interest. **Conclusion.** Based on the arguments put forth by Cotino Hueso and Llansó, it can be concluded that responsibility for this regulatory codification in the digital ecosystem lies directly with social media platform owners and with individual users who violate the rules, rather than with the capabilities of AI itself. Delegating judgment to a binary algorithm creates opacity and a form of "hygienic cleansing" that sacrifices democratic pluralism by subordinating fundamental rights to a logic of neutralizing linguistic harm, devoid of legal hermeneutics.
Keywords: Algorithmic content moderation, Artificial intelligence, ECtHR, Freedom of expression, Hate speech, Intermediary liability, Inter-American Court of Human Rights, Legal argumentation

Introducción

Las sociedades contemporáneas experimentan una mutación estructural en el ejercicio de la libertad de expresión, derivada del tránsito hacia ecosistemas digitales dominados por infraestructuras tecnológicas complejas. En este escenario, la Inteligencia Artificial (IA) ha dejado de ser una herramienta accesoria para constituirse como un actor normativo de facto y un intermediario central con facultades cuasi-regulatorias. Este fenómeno desplaza la regulación estatal hacia actores privados transnacionales que, mediante algoritmos escalables, jerarquizan o eliminan contenidos priorizando criterios de eficiencia sobre el rigor jurídico.

Esta realidad agudiza la tensión dialéctica entre la protección de la libertad de expresión y la salvaguarda de

la dignidad humana frente al discurso de odio. La gestión algorítmica, lejos de ser neutra, opera bajo modelos probabilísticos que a menudo derivan en una suerte de censura previa y en el fenómeno del *overblocking*. Como advierte Llansó, la carencia de comprensión contextual en estos sistemas automatizados erosiona el espacio del disenso y genera un vacío en la rendición de cuentas, desafiando la legitimidad democrática en la era digital.

Ante este panorama, el análisis de los sistemas regionales de protección de derechos humanos resulta estratégico. Mientras la jurisprudencia de la Corte Interamericana de Derechos Humanos (Corte IDH) mantiene una visión expansiva de la libertad de expresión —producto de su memoria histórica de resistencia al autoritarismo—, el modelo del Tribunal Europeo de Derechos Humanos (TEDH) prioriza la dignidad humana y la responsabilidad de los intermediarios, condicionado por su experiencia frente al totalitarismo.

El presente artículo analiza críticamente cómo la IA transforma la moderación del discurso de odio y fractura el sistema regulatorio clásico. Bajo un enfoque cualitativo y comparado, se integran las doctrinas de autores como Cotino Hueso, Echeveste y Presno Linera, junto con la jurisprudencia de los tribunales ya referidos. El objetivo es pormenorizar este fenómeno desde un plano jurídico-fundacional, el cual resulta indispensable para llamar la atención del lector sobre cómo las categorías normativas tradicionales son superadas por una gobernanza técnica que avanza de forma exponencialmente más rápida que la legislación civil.

Para delimitar de forma rigurosa este objeto de estudio, el análisis separa metodológicamente tres dimensiones operativas: el sujeto usuario que emplea la tecnología para desarrollar y amplificar discursos perjudiciales (como las fake news), los procesos automatizados desplegados en las redes sociales para detectar contenidos prohibidos, y las normativas internacionales a las que estas situaciones se deben acoger.

Bajo esta premisa, la hipótesis central sostiene que la Inteligencia Artificial por sí misma no es norma sino un instrumento técnico; por lo tanto, la verdadera responsabilidad de esta codificación regulatoria y su impacto directo sobre la libertad de expresión recae de forma exclusiva en las directrices fijadas por los dueños de las plataformas de redes sociales y en los perfiles individuales que violentan dichas pautas, más que en la capacidad mecánica de la tecnología.

Métodos

El presente estudio se desarrolló bajo un enfoque cualitativo con un diseño dogmático-jurídico y de análisis comparado. La investigación se centró en evaluar las líneas jurisprudenciales de la Corte Interamericana de Derechos Humanos y el Tribunal Europeo de Derechos Humanos respecto a la moderación algorítmica del discurso de odio.

Para ello, se seleccionaron sentencias hito, emitidas por ambos tribunales de protección regional, priorizando aquellos fallos que abordan la responsabilidad de los intermediarios digitales y las restricciones legítimas a la libertad de expresión en el ecosistema de las redes sociales. La técnica de recolección de información consistió en el análisis documental exhaustivo de los instrumentos convencionales, las opiniones consultivas y la jurisprudencia relevante.

El procesamiento de los datos se estructuró a partir de una matriz categorial fundamentada en las tesis doctrinales de Cotino Hueso y Llansó, contrastando los criterios materiales de incitación, daño y riesgo inminente frente a los efectos del sobre bloqueo algorítmico (*overblocking*), lo cual permitió contrastar la operatividad del constitucionalismo preventivo europeo con el modelo garantista expansivo interamericano.

Resultados

El marco jurídico del odio en el derecho internacional

La libertad de expresión como fundamento estructural del orden democrático

El derecho internacional contemporáneo reconoce la libertad de expresión no solo como un derecho subjetivo individual, sino como un pilar institucional del sistema democrático. Su función trasciende la autonomía comunicativa, pues constituye el mecanismo esencial para el debate público, la fiscalización del poder y la consolidación del pluralismo político. Esta doble dimensión —individual y colectiva— es recogida por la Convención Americana sobre Derechos Humanos y el Convenio Europeo de Derechos Humanos, instrumentos que la consagran como presupuesto indispensable para una democracia sustantiva.

Desde el constitucionalismo democrático, la legitimidad del poder político se vincula estrechamente con la protección de la libertad de expresión. Esta garantía no se limita a discursos socialmente aceptados, sino que ampara expresiones disidentes o perturbadoras del consenso mayoritario. No obstante, esta protección estructural no es absoluta; su ejercicio conlleva límites intrínsecos cuya delimitación técnica representa uno de los desafíos más complejos de la dogmática jurídica actual.

El test tripartito como criterio de restricción y su crisis en el entorno digital

El Derecho internacional, a través de la Observación General n.º 34 del Comité de Derechos Humanos (2011), ha consolidado el "test tripartito" como el estándar imperativo para evaluar la validez de las restricciones a la libertad de expresión, conforme al artículo 19 del PIDCP. Este instrumento exige la concurrencia simultánea de tres requisitos: legalidad (norma clara, precisa y accesible), finalidad legítima (protección de derechos de terceros, seguridad nacional, orden público, salud o moral pública)

y necesidad y proporcionalidad (la medida debe ser indispensable y la menos restrictiva en una sociedad democrática).

La aplicación práctica de este estándar enfrenta desafíos disruptivos en el ecosistema digital. Garay (2021) advierte que la moderación de contenidos por actores privados carece de supervisión judicial y procesos deliberativos, lo que desnaturaliza el rigor del test. En esta línea, Cotino Hueso (2017) señala que la imprecisión terminológica en el entorno virtual genera inseguridad jurídica y autocensura, fenómeno que Presno Linera (2010) identifica como el chilling effect o efecto desalentador, derivado de leyes con conceptos jurídicos indeterminados.

La proporcionalidad surge como el elemento más vulnerable en internet. Barata y Echeveste (2021) sostienen que las medidas actuales suelen ignorar la búsqueda del medio menos lesivo. Al respecto, Llansó (2015) argumenta que el bloqueo total de sitios o la eliminación automatizada constituyen medidas desproporcionadas que lesionan la infraestructura del discurso público.

En consecuencia, la validez de cualquier restricción debe interpretarse bajo el principio in dubio pro libertate; el incumplimiento de cualquiera de los parámetros del test debe resolverse en favor del ejercicio del derecho, reafirmando que la libertad de expresión es la regla y la limitación, la excepción.

La construcción jurídica del discurso de odio: entre la estabilidad de las reglas y la ductilidad de los principios

El ordenamiento internacional carece de una definición unívoca de "discurso de odio", operando como una categoría jurídica indeterminada cuyo contenido se supedita a contextos históricos y culturales. No obstante, doctrinas como las de Waldron (2012), Sunstein (2021) y el Plan de Acción de Rabat coinciden en una distinción funcional: la expresión ofensiva —protegida por el pluralismo— frente a la apología que incita directamente a la violencia o discriminación, objeto de prohibición técnica.

Esta dicotomía se analiza bajo dos paradigmas en tensión que definen el constitucionalismo contemporáneo (Prieto Sanchís, 2003):

1. El modelo de reglas y constitucionalismo de resistencia: Autores como Schauer (2004), Habermas (2010) y Ferrajoli (2011) defienden la "esfera de lo indecidible", donde los derechos fundamentales constituyen normas deónticas que no deben ser sometidas a balances de utilidad. Bajo este enfoque, Presno Linera (2014) y Ansuátegui Roig (2010) abogan por la "delimitación" del contenido del derecho en lugar de la ponderación judicial, evitando que el activismo de los jueces erosione el núcleo de las libertades mediante criterios subjetivos. Para esta corriente, la progresividad de los derechos depende de su fijeza como límites infranqueables al poder.

2. El modelo de principios o teoría de la ponderación: Preponderante en el Sistema Interamericano y desarrollado por Robert Alexy (2019), este modelo concibe los derechos como "mandatos de optimización" que requieren un examen de proporcionalidad en tres niveles: idoneidad, necesidad y proporcionalidad en sentido estricto. Esta visión es respaldada por la tesis del "derecho dúctil" de Zagrebelsky (2011) y la estructura procedimental de Atienza (2017), quienes sostienen que los principios no pueden jerarquizarse de forma absoluta, obligando al juez a un ejercicio de equilibrio constante. Bernal Pulido (2014) consolida esta postura al proponer la proporcionalidad como el criterio de legitimidad técnica que otorga objetividad a la restricción.

Como síntesis de esta tensión, emerge el modelo intermedio de Jeremy Waldron (2005). Desde una dimensión institucional, Waldron propugna por el debate parlamentario como la vía legítima para resolver desacuerdos sobre derechos, rechazando la exclusividad de la imposición judicial. En el plano sustantivo, su tesis justifica la restricción del discurso de odio en la protección de la "seguridad pública" y el resguardo de la dignidad, entendida como el estatus social básico de los sujetos.

El daño jurídico no reside en la afectación de sentimientos, sino en el socavamiento de la confianza de los grupos vulnerables como miembros de pleno derecho en la comunidad. Una vez delimitados estos fundamentos teóricos y el "test tripartito" como marco formal, corresponde analizar los requisitos de fondo que legitiman la sanción en el entorno digital.

Criterios materiales de intervención: incitación, daño y riesgo

En el marco del Derecho Internacional de los Derechos Humanos, la restricción de la libertad de expresión posee un carácter excepcional y taxativo, orientado a prevenir el uso de la regulación como instrumento de censura arbitraria. Más allá del cumplimiento formal del test tripartito, la legitimidad de cualquier intervención estatal o privada requiere la concurrencia de criterios materiales específicos.

Estos parámetros permiten distinguir las expresiones polémicas o hirientes —amparadas por el pluralismo democrático— de aquellas que representan una amenaza real para la integridad de terceros y la paz social. En el contexto de la inmediatez de las redes sociales, la identificación de estos límites es crítica: deben interpretarse no como meras reglas de restricción, sino como garantías de la libertad misma.

Una aplicación laxa de estos criterios bajo el pretexto de protección conlleva el riesgo de silenciar voces disidentes, erosionando la diversidad de pensamiento que fundamenta la vida democrática. Por tanto, la intervención sobre el discurso requiere un escrutinio riguroso que evite la transición de la protección legítima hacia la censura sistémica.

Configuración de los presupuestos sustantivos: incitación, daño y riesgo

Conforme a los estándares del Plan de Acción de Rabat, el artículo 20 del PIDCP, la jurisprudencia internacional identifica tres presupuestos sustantivos: la incitación directa, el daño verificable y el riesgo inminente.

1. Incitación directa e intencionalidad: De acuerdo con la ACNUDH (2020), no basta con que un discurso sea impopular; se requiere una intención deliberada de promover hostilidad. El contenido debe poseer un carácter exhortativo, constituyendo un "llamado a la acción" que inste a la discriminación o violencia contra grupos específicos. Se excluyen expresiones negligentes o imprudentes, centrándose en aquellas donde la voluntad de causar daño y la capacidad de movilización son evidentes.
2. Daño verificable y nexo causal: La intervención estatal no puede sustentarse en la reprobación moral o la ofensa subjetiva. Debe demostrarse un nexo causal entre la expresión y un impacto real, o una probabilidad altamente valorable de vulneración de derechos (Naciones Unidas, 2013). La Corte IDH (2009) sostiene que el pluralismo exige proteger incluso informaciones que puedan "chocar o perturbar"; por tanto, el daño debe ser objetivo, traduciéndose en actos de discriminación efectiva, ataques a la integridad física o supresión de derechos civiles.
3. Riesgo inminente y contexto social: Este criterio proscribe la "censura por sospecha" basada en peligros hipotéticos.

El riesgo debe ser concreto y próximo. Según la OEA (2017), esto exige evaluar el contexto social: el impacto de una proclama varía drásticamente según la estabilidad democrática o la polarización del entorno. En el ecosistema digital, aunque la viralización acelera la movilización, la autoridad debe verificar si el emisor posee la audiencia y autoridad suficiente para que su discurso desencadene acciones violentas en el corto plazo (ACNUDH, 2020).

Aplicabilidad de los criterios materiales en los sistemas regionales de protección

Los criterios de incitación, daño y riesgo no se limitan al ámbito del sistema universal, sino que constituyen el núcleo del control de convencionalidad en los sistemas Interamericano y Europeo de Derechos Humanos. Aunque ambos regímenes convergen en el estándar de necesidad en una sociedad democrática, han desarrollado metodologías diferenciadas para evaluar la legitimidad de las restricciones.

Mientras el Sistema Interamericano prioriza la dimensión social de la libertad de expresión como garantía contra el autoritarismo, el Sistema Europeo otorga un margen de apreciación más amplio a los Estados para proteger la dignidad humana y el orden público. Esta divergencia procedimental no altera la obligatoriedad de los presupuestos materiales, pero condiciona su aplicación práctica frente a discursos que desafían la cohesión social y la estabilidad institucional en sus respectivos contextos regionales.

Divergencias funcionales: el Sistema Interamericano frente al europeo

La aplicación de los criterios materiales de restricción varía según la trayectoria histórica y dogmática de cada sistema regional. El Sistema Interamericano, condicionado por un pasado de regímenes autoritarios, mantiene un estándar de protección expansivo. La jurisprudencia de la Corte IDH (2009) establece un umbral de restricción sumamente elevado, priorizando el régimen de responsabilidades ulteriores sobre cualquier forma de censura previa. En este contexto, el criterio de riesgo inminente se aplica con rigorismo excepcional, exigiendo la prueba fehaciente de una amenaza real y directa a la vida o integridad para legitimar la intervención sobre el discurso.

En contraste, el Sistema Europeo opera bajo la doctrina del margen de apreciación nacional, que otorga a los Estados una mayor potestad para restringir expresiones que afecten la moral pública o la paz social, particularmente en ámbitos religiosos o de memoria histórica. No obstante, en el ámbito del discurso político, el TEDH (2019) converge con los estándares de Rabat, aplicando un test de proporcionalidad estricto. Este análisis distingue entre la provocación legítima en debates de interés público y la incitación efectiva al odio, garantizando que la limitación no erosione el pluralismo político esencial de la región.

La gobernanza algorítmica: la transformación del discurso jurídico en el entorno digital.

El entorno digital no ha generado un vacío legal, sino lo que algunos teóricos denominan una "hiperregulación técnica". A diferencia del entorno analógico, en el digital la norma no solo emana del Estado, sino de la arquitectura del código. "La arquitectura es la ley" (*Code is law*), como señalaba Lawrence Lessig (2006). En el ecosistema digital, la norma se intensifica porque la sanción no es posterior a la conducta, sino preventiva: el sistema simplemente impide que la acción ocurra.

En esa misma línea de pensamiento, Garay (2016) identifica cuatro factores determinantes que transforman el alcance del mensaje y, por ende, su peligrosidad jurídica: la escalabilidad, que permite la difusión masiva y casi instantánea a millones de usuarios; la persistencia, que garantiza la disponibilidad indefinida de la información; la descontextualización, que fragmenta el sentido original del mensaje; y el anonimato relativo, que reduce la barrera de la responsabilidad social y legal del emisor. La convergencia de estos elementos potencia de forma exponencial el impacto del discurso de odio y la desinformación, generando una presión sin precedentes sobre los Estados. Esta realidad obliga a los sistemas jurídicos a transitar desde modelos de responsabilidad reactiva hacia respuestas normativas más robustas y preventivas que intenten mitigar los riesgos inherentes a una esfera pública tecnológicamente mediada.

La transición hacia una gobernanza digital delegada ha consolidado lo que Llansó (2020) denomina un nuevo paradigma en la gestión del discurso: la moderación automatizada a gran escala. Bajo este modelo, los sistemas de Inteligencia Artificial operan estrictamente mediante patrones estadísticos y una simplificación del lenguaje que ignora la complejidad de las interacciones humanas. Al no poseer una capacidad cognitiva para comprender intenciones o matices pragmáticos, estas herramientas tienden a reproducir sesgos sistémicos presentes en sus conjuntos de entrenamiento. El resultado es una "censura algorítmica" que carece de motivación jurídica y opera bajo una profunda opacidad decisional, eliminando la posibilidad de una contextualización profunda o un ejercicio de contradicción inmediata.

Desde la perspectiva del derecho internacional y constitucional, este fenómeno desafía directamente los principios de legalidad y debido proceso, al sustituir el juicio humano ponderado por una ejecución automatizada, técnica y, a menudo, arbitraria.

Como efecto de lo antes descrito en este análisis, encontramos que uno de los efectos más disruptivos del ecosistema contemporáneo, es la transferencia de la soberanía regulatoria desde el Estado hacia actores privados transnacionales, consolidando lo que diversos autores entre ellos, Celeste, E. (2022) denominan una "soberanía digital de facto". En este escenario, las grandes plataformas tecnológicas han dejado de ser meros intermediarios para convertirse en autoridades paraestatales que definen reglas de conducta, interpretan normas de manera unilateral y ejecutan sanciones —como la suspensión de perfiles o el bloqueo de contenidos— con una celeridad que el sistema judicial no puede igualar. Al decidir la visibilidad de los discursos mediante algoritmos, estas empresas ejercen una gobernanza privada que carece de legitimidad democrática directa, operando al margen de los mecanismos de control constitucional y desplazando la seguridad jurídica por términos y condiciones de servicio que supeditan los derechos fundamentales a intereses comerciales.

Por su parte, Cotino Hueso (2019) denomina esta “privatización normativa” como un proceso donde las empresas tecnológicas dejan de ser meras espectadoras, para incurrir en actividades de regulación informal y omnipresencia del discurso público global. Al dictar sus propios estándares de convivencia digital, estas corporaciones instauran un orden paralegal que debilita drásticamente el control judicial, al trasladar la resolución de conflictos desde los tribunales públicos hacia sistemas internos de apelación corporativa.

Esta dinámica no solo compromete la transparencia de los procesos sancionatorios, sino que diluye la responsabilidad política, pues el Estado se ve incapaz de garantizar los derechos fundamentales frente a decisiones algorítmicas tomadas en jurisdicciones extranjeras. En definitiva, la privatización normativa desplaza el interés general por políticas de uso privadas, dejando al ciudadano en un estado de vulnerabilidad procesal ante el nuevo poder regulador del siglo XXI.

El Derecho Internacional, ante esta complejidad, ha comenzado a articular respuestas mediante documentos orientadores que buscan someter la técnica al imperio de los derechos humanos. En este sentido, el Consejo de Europa (2020) ha establecido estándares imperativos que exigen la realización de evaluaciones de impacto en los derechos fundamentales, la garantía de una transparencia algorítmica efectiva y, primordialmente, el mantenimiento del control humano sobre las decisiones automatizadas.

En sintonía con este esfuerzo, la Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos (RELE-CIDH, 2023) ha emitido alertas críticas sobre los riesgos de la censura privada y la ausencia de un debido proceso digital. Según este organismo, la opacidad de los sistemas de moderación no solo vulnera las garantías individuales, sino que lesiona el pluralismo democrático al uniformar el discurso bajo criterios corporativos.

Como se puede observar, estos instrumentos evidencian una convergencia jurídica global: la urgencia de regular la Inteligencia Artificial no desde una perspectiva puramente comercial, sino desde un enfoque de derechos humanos que devuelva al ciudadano las garantías arrebatadas por la automatización.

El estándar interamericano: protección reforzada y dimensión estructural

Fundamentos normativos y la tesis de la doble dimensión

El estándar interamericano trasciende la concepción clásica del derecho subjetivo para consolidarse como una garantía estructural del orden democrático. A partir de la Opinión Consultiva OC-5/85, la Corte Interamericana de Derechos Humanos (1985) estableció la naturaleza dual de este derecho: una dimensión individual, referida a la autonomía para expresar ideas, y una dimensión social, que garantiza el derecho de la colectividad a recibir información diversa. Esta arquitectura jurídica redefine la libertad de expresión como un bien jurídico tanto privado como colectivo, donde cualquier restricción indebida genera un daño sistémico que erosiona el pluralismo (Echeveste, 2015).

Desde una perspectiva filosófica, este enfoque se sustenta en la teoría del "mercado abierto de ideas", indispensable para el ejercicio del control ciudadano sobre el poder. Como sostiene Bertoni (2012), la tradición regional marcada por el autoritarismo ha obligado al SIDH a blindar el discurso público como una "piedra angular" de la democracia. Bajo esta premisa, el sistema manifiesta una resistencia categórica a legitimar limitaciones basadas en conceptos jurídicos indeterminados —como la "ofensa" o la "armonía cultural"—, entendiendo que el sacrificio de la expresión en aras de una supuesta paz social suele encubrir mecanismos de control hegemónico que asfixian la vitalidad democrática y la fiscalización de la arbitrariedad.

Jurisprudencia fundacional: construcción progresiva del estándar interamericano frente al entorno digital

La evolución del estándar interamericano responde a una construcción progresiva frente a las tensiones entre el poder y el discurso. Esta trayectoria permite hoy cuestionar la legalidad de la moderación algorítmica mediante principios decantados en fallos históricos de la Corte IDH.

El cimiento de este régimen reside en la proscripción de la censura previa (Olmedo Bustos vs. Chile, 2001), principio que invalida el uso de algoritmos de "sensibilidad" como barreras preventivas al flujo informativo. Complementariamente, el hito del acceso a la información pública (Claude Reyes vs. Chile, 2006) fundamenta la exigencia de transparencia algorítmica sobre la visibilidad de contenidos. Por su parte, el deber de especial cuidado desarrollado en el caso Ríos y otros vs. Venezuela (2009) proyecta la responsabilidad de las plataformas en la protección de voces críticas frente al acoso digital (trolling).

La protección reforzada del discurso de interés público fue ratificada en el caso Mémoli vs. Argentina (2013), subrayando que la simplificación técnica de la IA no justifica sanciones automáticas descontextualizadas. Finalmente, el caso Álvarez Ramos vs. Venezuela (2019) reafirma que el uso de mecanismos sancionatorios —estatales o privados— para silenciar críticas es incompatible con la Convención Americana.

En síntesis, esta jurisprudencia constituye un "bloque de resistencia" (Bertoni, 2012) donde la protección del discurso de interés público se erige como una norma de *ius cogens* regional (García Ramírez, 2006; Cançado Trindade, 2018). En consecuencia, la "privatización normativa" (Cotino Hueso, 2019) no implica anomia; cualquier intervención en el entorno digital debe someterse estrictamente al test de proporcionalidad y a las garantías del debido proceso (Art. 8 CADH), asegurando que la tecnología no devenga en un instrumento de regresión democrática.

El discurso político y el tratamiento del odio en el Sistema Interamericano (SIDH)

El estándar interamericano otorga al discurso sobre asuntos de interés público una presunción de primacía, constituyendo el "núcleo duro" del artículo 13 de la Convención Americana. Esta protección reforzada exige que el umbral de tolerancia frente a la crítica sea mayor cuando el sujeto afectado es un funcionario o institución estatal. Como sostiene García Ramírez (2006), el control ciudadano es una condición sine qua non para la transparencia, lo que somete cualquier restricción en este ámbito a un escrutinio judicial estrictísimo. Este estándar colisiona con la "armonía digital" que priorizan los algoritmos de las plataformas, los cuales, al simplificar el lenguaje, corren el riesgo de vaciar de contenido la crítica política o satírica (Botero Marino, 2013).

En cuanto al discurso de odio, el SIDH aplica un criterio de excepcionalidad y un enfoque restrictivo basado en el artículo 13.5 de la Convención. Para que una expresión sea sancionable, no basta con que sea ofensiva o perturbadora; debe constituir una incitación directa a la violencia o a la acción ilegal, demostrando un nexo causal claro con un riesgo real e inminente (Lanza, 2017). Esta interpretación técnica limita la facultad de la IA para etiquetar como "odio" expresiones vehementes que no alcanzan dicho umbral de peligrosidad.

Finalmente, el sistema blinda el denominado "discurso incómodo", protegiendo incluso informaciones que "ofenden, resultan ingratas o perturban al Estado" (Corte IDH, Caso Olmedo Bustos, 2001). Este principio actúa como un antídoto frente a la "moralidad algorítmica" de las corporaciones tecnológicas. Como apunta Fiss (2004), la vitalidad democrática requiere proteger expresiones que, aunque socialmente impopulares, son esenciales para el control del poder. Así, la RELE-CIDH (2023) advierte sobre la instrumentalización de definiciones vagas de "odio" como pretexto para la persecución ideológica, reafirmando que la protección de la armonía social no debe suprimir el pluralismo político.

Gobernanza algorítmica en el Sistema Interamericano: intermediarios y el fenómeno del overblocking

El ecosistema digital ha transformado a las plataformas tecnológicas en los nuevos "porteros" (gatekeepers) del flujo informativo global. Bajo el estándar interamericano, estos intermediarios no son receptáculos pasivos, sino actores que ejercen una función de gobernanza con impacto directo en la dimensión social de la libertad de expresión (RELE-CIDH, 2017). No obstante, el sistema defiende el principio de responsabilidad limitada, advirtiendo que obligar a las plataformas a un monitoreo preventivo universal las convertiría en "censores privados" por aversión al riesgo legal o económico.

Esta dinámica subyace al fenómeno del *overblocking* o sobre bloqueo algorítmico. Este ocurre cuando los sistemas de IA, programados bajo lógicas de eficiencia estadística, eliminan masivamente contenidos legítimos, satíricos o periodísticos al confundirlos con discursos proscritos. Como señala Llansó (2020), la IA carece de la capacidad cognitiva para realizar el juicio de ponderación exigido por la jurisprudencia interamericana (Caso Mémoli vs. Argentina, 2013). El resultado es una purga del debate público donde el "algoritmo del error" prefiere la censura por exceso, vulnerando los principios de necesidad y proporcionalidad.

Frente a la opacidad decisional, el enfoque regional exige la transición hacia un debido proceso algorítmico. Siguiendo las directrices de la RELE-CIDH (2023) y Kaye (2019), la legitimidad de la moderación automatizada depende de garantías mínimas: notificación clara de la infracción, mecanismos de revisión humana y transparencia técnica mediante auditorías de sesgos. En última instancia, el acceso a las plataformas digitales constituye hoy el acceso al ejercicio de la ciudadanía, lo

que obliga a someter la "justicia privada" corporativa al imperio de los derechos humanos.

El estándar europeo: hacia el endurecimiento de la responsabilidad

Fundamentos del Artículo 10 CEDH y el margen de apreciación nacional

A diferencia de la tradición interamericana, el modelo europeo, anclado en el artículo 10 del Convenio Europeo de Derechos Humanos (CEDH), se define por una concepción de "democracia militante" (Grimm, 2009). Esta doctrina permite restricciones más amplias para proteger el orden democrático y los derechos de terceros, bajo la premisa de que la libertad de expresión conlleva "deberes y responsabilidades" (Caso Handyside vs. Reino Unido, 1976). Este enfoque funcional justifica la imposición de límites proporcionales cuando el ejercicio del derecho amenaza la dignidad humana o la paz social (Barendt, 2005).

La operatividad de estas restricciones se articula mediante el margen de apreciación nacional, que otorga a los Estados cierta discrecionalidad para evaluar la "necesidad social imperiosa" de limitar un discurso. No obstante, en la era de la Inteligencia Artificial, esta doctrina enfrenta el riesgo de una "anomia algorítmica" (Presno Linera, 2022). Bajo los principios de reserva de ley y proscripción de la arbitrariedad, el margen de apreciación no puede derivar en una transferencia de soberanía hacia sistemas automatizados; por el contrario, exige una transparencia radical y un “debido proceso tecnológico” que garantice que la moderación no sea arbitraria, sino auditable y fundamentada en la ley (Voorhoof, 2014).

El Artículo 17 CEDH y la proscripción del abuso del derecho

Una característica distintiva del sistema europeo es la cláusula de salvaguarda del artículo 17 del CEDH, diseñada para impedir que las libertades reconocidas sean utilizadas para destruir los fundamentos de la democracia. Esta "cláusula de exclusión" retira la protección del Convenio a discursos de naturaleza totalitaria o de odio extremo —como el negacionismo del Holocausto—, los cuales, según la jurisprudencia del TEDH, no ameritan siquiera el examen de proporcionalidad del artículo 10.2 (García Roca, 2010).

En el contexto de la algoritmocracia, la aplicación de esta cláusula adquiere una relevancia crítica. Presno Linera (2022) advierte que la calificación de una expresión como "abusiva" requiere un juicio de valor eminentemente humano que la IA no puede replicar. El riesgo radica en que las plataformas privadas utilicen la retórica del "abuso del derecho" para normalizar una privatización de la excepción, eliminando discursos incómodos bajo la apariencia de proteger la democracia. Por tanto, la exclusión del artículo 17 debe ser restrictiva y estar sujeta a un control jurisdiccional riguroso, evitando que una herramienta de defensa democrática se convierta, mediante ejecuciones algorítmicas opacas, en una vía para la anulación del pluralismo político (Barendt, 2005).

Evolución jurisprudencial del TEDH: responsabilidad de intermediarios y diligencia digital

La construcción del estándar europeo sobre la responsabilidad de los intermediarios digitales presenta una trayectoria pendular que busca equilibrar la libertad de información con la protección de la dignidad humana. Esta evolución permite identificar los límites del margen de apreciación frente a la gestión automatizada de contenidos.

El hito fundacional se establece en el caso Delfi AS c. Estonia (2015), donde la Gran Sala determinó que los portales comerciales poseen un "deber de vigilancia activa" y deben retirar discursos de odio manifiesto sin demora. No obstante, este criterio fue matizado en MTE e Index.hu c. Hungría (2016), al establecer que la responsabilidad objetiva vulnera el artículo 10 del CEDH si los contenidos no constituyen incitación evidente a la violencia, protegiendo así a los intermediarios frente al riesgo de *overblocking*.

Posteriormente, el caso Beizaras y Levickas c. Lituania (2020) reforzó la "obligación positiva de protección" del Estado para investigar el odio privado en redes sociales, asegurando que el entorno digital no sea un espacio de impunidad. Por su parte, la sentencia Perinçek c. Suiza (2015) subrayó la importancia del contexto en el discurso histórico y político, reafirmando que la simplificación algorítmica no puede sustituir el juicio humano en temas de interés público.

Finalmente, el caso Sánchez c. Francia (2023) introdujo un giro disruptivo al desplazar la responsabilidad penal hacia el titular del perfil (político) por no moderar comentarios de terceros en su muro, consolidando un deber de vigilancia personal en el ecosistema digital.

En síntesis, la jurisprudencia del TEDH revela un tránsito hacia lo que Presno Linera (2022) denomina la "constitucionalización de la moderación". Este modelo de corresponsabilidad diligente exige una ponderación casuística que considere el contexto y el carácter del espacio digital. Como advierte Voorhoof (2023), el sistema europeo ha erigido un estándar que obliga a los intermediarios a actuar como garantes de la dignidad humana, siempre bajo el tamiz de una "necesidad social imperiosa" que evite que el control preventivo devenga en una anulación del pluralismo democrático.

Doctrina crítica y gobernanza algorítmica: el riesgo del "endurecimiento" y el *overblocking*

La reciente deriva jurisprudencial del TEDH, desde Delfi hasta Sánchez c. Francia, ha generado una preocupación académica sobre el "endurecimiento peligroso" de las obligaciones digitales. Cotino Hueso (2023) advierte que la expansión de la responsabilidad indirecta hacia plataformas y usuarios administradores crea un incentivo perverso al *overblocking*.

Ante el temor a sanciones, los intermediarios optan por una "privatización preventiva" del control discursivo, transformándose en censores privados que anticipan juicios de valor de competencia judicial.

Presno Linera (2022) añade que imponer deberes de cuidado excesivos a sistemas automatizados —carentes de sensibilidad contextual— asfixia el pluralismo político en nombre de la seguridad jurídica, institucionalizando una vigilancia que debilita la esfera pública democrática.

Frente a esta "anomia digital", el modelo europeo ha transitado hacia una gobernanza basada en el principio de responsabilidad proactiva. Un hito fundamental es el Reglamento (UE) 2024/1689 (Ley de IA), que clasifica a los sistemas de moderación de contenidos como tecnologías de alto riesgo. Este marco normativo exige que los algoritmos abandonen su naturaleza de "cajas negras" para cumplir con estándares de transparencia técnica y supervisión humana, proscribiendo decisiones automatizadas sin base lógica auditable (Presno Linera, 2022).

El mayor desafío de esta gobernanza sigue siendo la "censura privada algorítmica", donde la eficiencia estadística prevalece sobre la justicia del caso concreto. No obstante, el modelo propone un constitucionalismo digital que traslada las garantías del proceso equitativo (Art. 6 CEDH) al entorno virtual. Según Barendt (2005) y Voorhoof (2023), la gobernanza algorítmica solo adquiere legitimidad si permite la contradicción y la revisión humana, exigiendo a las plataformas una diligencia debida que proteja la integridad de la esfera pública frente a la simplificación tecnológica.

Discusión

Análisis comparado: convergencias y divergencias en la gobernanza del discurso en la era algorítmica

La confrontación dialéctica entre los sistemas Interamericano y europeo revela un panorama jurídico donde la tecnología actúa como catalizador de nuevas tensiones constitucionales. Más allá de sus diferencias de origen, ambos modelos operan sobre una arquitectura de principios que sitúa a la dignidad humana y al sistema democrático como límites infranqueables a la técnica (principio pro homine).

Convergencias axiológicas: la centralidad de los derechos frente al algoritmo

Existe un consenso transatlántico que eleva la libertad de expresión a la categoría de pilar estructural de la democracia. Desde la OC-5/85 de la Corte IDH hasta el caso Handyside vs. Reino Unido del TEDH, se reconoce que este derecho protege incluso las ideas que "ofenden, chocan o perturban", siendo una herramienta primaria para el control del poder (García Ramírez, 2006). Asimismo, ambos sistemas coinciden en la no absolutez del derecho, supeditando cualquier restricción al cumplimiento estricto del test tripartito: legalidad, finalidad legítima y proporcionalidad.

En la lucha contra el discurso discriminatorio, el sistema europeo (Art. 17 CEDH) y el interamericano (Art. 13.5 CADH) convergen en que la apología del odio no constituye una categoría protegida. Esta zona de coincidencia ética exige que las plataformas asuman una responsabilidad proactiva en la protección de grupos vulnerables, abandonando la visión de la red como un espacio de anomia.

Como sostiene Presno Linera (2022), tanto en Estrasburgo como en San José se exige que la moderación de contenidos respete la reserva de ley y proscriba la arbitrariedad técnica, identificando al *overblocking* como una amenaza sistemática al pluralismo (Cotino Hueso, 2023).

Convergencia funcional: debida diligencia y protección del discurso político

A nivel procedimental, se observa una transversalidad de garantías orientada a la creación de un estándar de debida diligencia algorítmica. Ambos modelos exigen que la moderación abandone la opacidad técnica para someterse a auditorías, transparencia en el entrenamiento de la IA y mecanismos de apelación efectiva. Este marco de "constitucionalismo digital" busca someter el poder de las corporaciones tecnológicas al control jurisdiccional efectivo.

Finalmente, ambos sistemas otorgan una protección reforzada al discurso político. El TEDH, en el caso MTE e Index.hu c. Hungría (2016), y la Corte IDH, en el caso Álvarez Ramos c. Venezuela (2019), coinciden en que el umbral de tolerancia frente a la crítica debe ser mayor cuando se trata de funcionarios públicos o temas de relevancia social. Esta "presunción de primacía" limita la capacidad de la IA para ejecutar limpiezas reputacionales automatizadas, garantizando que el entorno digital permanezca como un espacio de deliberación democrática y no de conformidad manufacturada (RELE-CIDH, 2023).

Divergencias operativas y modelos de gobernanza ante la algoritmocracia

A pesar de las convergencias axiológicas, existen fracturas doctrinales que definen respuestas jurídicas diferenciadas frente al fenómeno tecnológico:

Escrutinio estricto frente al margen de apreciación nacional

La metodología de restricción revela la primera gran divergencia. El Sistema Interamericano, fiel a su tradición de resistencia al autoritarismo, mantiene un escrutinio estricto que prohíbe casi categóricamente la censura previa (Olmedo Bustos vs. Chile, 2001). En contraste, el Modelo Europeo emplea la doctrina del margen de apreciación nacional, otorgando a los Estados discrecionalidad para intervenir ante una "necesidad social imperiosa" (Handyside vs. Reino Unido, 1976).

Esta distinción es crítica: mientras América blinda el discurso político frente a la automatización, Europa permite una intervención algorítmica más enérgica para salvaguardar la dignidad humana, asumiendo el riesgo de una privatización normativa más profunda.

Umbrales de intervención y responsabilidad digital

El presupuesto habilitante para limitar el discurso también difiere. El SIDH adopta un modelo de libertad expansiva, exigiendo la prueba de incitación directa y riesgo inminente para validar cualquier restricción. Por el contrario, el TEDH se adscribe a un modelo de libertad condicionada por la cohesión social, utilizando el artículo 17 del CEDH para excluir discursos que atenten contra los valores democráticos.

Esta disparidad se traslada al rol de los intermediarios: el modelo interamericano rechaza la responsabilidad objetiva para evitar que las plataformas devengan en "censores privados". En cambio, el sistema europeo, desde Delfi AS hasta Sánchez c. Francia (2023), ha consolidado una responsabilidad indirecta que incentiva la supervisión activa. Como advierte Cotino Hueso (2023), este enfoque desplaza la función jurisdiccional hacia algoritmos de decisión privada obligados a anticipar riesgos legales.

Inteligencia Artificial: Constitucionalismo preventivo frente al expresivo

Finalmente, la respuesta ante la IA revela dos filosofías constitucionales. Europa apuesta por un constitucionalismo preventivo, materializado en el Reglamento (UE) 2024/1689 (AI Act), que regula estructuralmente los sistemas algorítmicos mediante la gestión de riesgos. El modelo interamericano se define como un constitucionalismo expresivo: ante la ausencia de un marco normativo similar, problematiza con mayor énfasis el impacto del *overblocking* en el disenso.

El ecosistema europeo genera incentivos hacia la eliminación masiva de contenidos por seguridad jurídica, por su parte, el sistema interamericano prioriza la protección del pluralismo frente al abuso estatal o corporativo, evidenciando, no obstante, un déficit en la gobernanza efectiva de las grandes plataformas tecnológicas.

Escenarios prospectivos en la gobernanza digital: ¿convergencia regulatoria o fragmentación asimétrica?

El futuro de la libertad de expresión en el ecosistema digital se debate entre la armonización global de estándares y el reforzamiento de las particularidades regionales. Al analizar la trayectoria dialéctica del Sistema Interamericano y el Modelo Europeo, se perfilan tres horizontes que determinarán la vigencia de los derechos fundamentales frente a la algoritmocracia:

Hacia un núcleo mínimo de constitucionalismo digital

Este escenario supone la convergencia de ambos sistemas en principios irrenunciables de control. Como sostiene Presno Linera (2022), la “transversalidad de garantías” forzará la consolidación de un núcleo duro compuesto por la transparencia algorítmica, el control humano efectivo y una proporcionalidad reforzada.

En este modelo, la supervisión judicial recupera su centralidad, impidiendo que la IA actúe como un tribunal de adjudicación privada. Tanto las directrices de la RELE-CIDH (2023) como el Reglamento (UE) 2024/1689 señalan un diálogo normativo donde la técnica se somete a una auditoría democrática para proscribir la arbitrariedad de la "caja negra".

La expansión del "Efecto Bruselas" y la presión preventiva

Un segundo horizonte se define por el "Efecto Bruselas" (Bradford, 2020), mediante el cual los estándares de la Unión Europea se expanden globalmente, induciendo al Sistema Interamericano a adoptar modelos preventivos. Esta tendencia, impulsada por la necesidad de combatir la desinformación, podría obligar al sistema regional a flexibilizar su histórica desconfianza hacia la responsabilidad de los intermediarios.

No obstante, Cotino Hueso (2023) advierte que este endurecimiento —validado en precedentes como Sánchez c. Francia (2023)— resultaría contraproducente en contextos latinoamericanos con instituciones debilitadas, donde la presión legal sobre las plataformas deriva frecuentemente en una censura privada preventiva que afecta desproporcionadamente a voces críticas.

Cristalización de paradigmas diferenciados de libertad

Finalmente, existe la posibilidad de una fragmentación doctrinal permanente basada en identidades filosóficas fundacionales. Europa se consolidaría como el paradigma regulador preventivo, priorizando la dignidad y la cohesión social; mientras que el Sistema Interamericano reafirmaría su rol como paradigma garantista expansivo, priorizando el pluralismo y el disenso. Como ha reiterado la Corte IDH (Álvarez Ramos vs. Venezuela, 2019), la libertad de expresión es la garantía existencial de la democracia.

Según Lanza (2017), esta visión debe resistir la tentación de una regulación anticipatoria que, en nombre de la eficiencia técnica o la paz social, termine por institucionalizar el silencio y reducir la vitalidad del debate público.

Conclusión

La gobernanza algorítmica instaura una soberanía digital de facto en las redes sociales que desplaza el paradigma clásico de protección de los derechos fundamentales hacia infraestructuras tecnológicas privadas. En este entorno, la inteligencia artificial opera estrictamente como un instrumento técnico, transfiriendo la verdadera responsabilidad convencional de la codificación censora a los dueños de las plataformas corporativas y a los perfiles individuales emisores.

Esta privatización normativa introduce una brecha crítica frente al test de proporcionalidad, pues la opacidad de los algoritmos y la indeterminación conceptual del

discurso de odio transmutan la restricción excepcional en una regla operativa preventiva y opaca, provocando un efecto desalentador (*chilling effect*) que sacrifica el pluralismo democrático.

Ante este escenario, el Sistema Interamericano defiende una doctrina de protección reforzada del disenso frente a regresiones autoritarias, aunque enfrenta un marcado déficit regulatorio frente a corporaciones transnacionales al depender principalmente de instrumentos de *soft law*. Por su parte, el modelo europeo avanza con un esquema de responsabilidad proactiva y diligencia debida a través de su Ley de Inteligencia Artificial; no obstante, el endurecimiento de las obligaciones digitales para los intermediarios delega funciones jurisdiccionales en actores comerciales, incentivando un perverso sobrebloqueo algorítmico (*overblocking*) que sustituye la ponderación humana por ejecuciones técnicas automatizadas.

La prospectiva de la gobernanza digital se halla en una encrucijada entre la expansión global del modelo preventivo europeo por el denominado "Efecto Bruselas" y el riesgo de una fragmentación regional que ignore la fragilidad institucional de contextos como el latinoamericano, donde la prioridad absoluta debe ser blindar el pluralismo frente al abuso estatal o corporativo. En consecuencia, la supervivencia del Estado de Derecho exige la consolidación de un constitucionalismo algorítmico que reconquiste la autoridad regulatoria sobre la esfera pública virtual, asegurando que las directrices corporativas privadas permanezcan supeditadas de manera vinculante a los principios de legalidad, transparencia y control judicial efectivo para proteger de forma irrestricta el pensamiento y el disenso de la población.

Conflictos de interés

La autora declara que no existe conflicto de interés de ningún tipo.

Financiamiento

Ninguno.

Consideraciones éticas

No aplica.

Uso de IA

Se declara el uso de herramientas de inteligencia artificial únicamente para apoyo gramatical y estructural, sin intervenir en el análisis ni en los resultados del estudio.

Referencias

Doctrina y artículos científicos

Abramovich, V., y Courtis, C. (2018). La protección de los derechos humanos en el sistema interamericano. Editores del Puerto.

Alexy, R. (2019). Teoría de los derechos fundamentales (E. Garzón Valdés, Trad.; 3.ª ed.). Centro de Estudios Políticos y Constitucionales. (Obra original publicada en 1986).

Ansuátegui Roig, F. J. (2010). Derechos fundamentales, valores y normas. Marcial Pons.

Atienza, M. (2017). El derecho como argumentación. Ariel.

Barata, J., y Echeveste, P. (2021). Libertad de expresión y entornos digitales: El desafío de la proporcionalidad. Editorial Dykinson.

Barendt, E. (2005). Freedom of speech (2.ª ed.). Oxford University Press.

Bernal Pulido, C. (2014). El principio de proporcionalidad y los derechos fundamentales (4.ª ed.). Universidad Externado de Colombia.

Bertoni, E. (2012). Prevenir la censura: El Sistema Interamericano de Derechos Humanos y la libertad de expresión. Editorial Universidad de Palermo.

Botero Marino, C. (2013). El estándar interamericano de libertad de expresión y los desafíos de la era digital. En Derecho y Tecnología. Universidad de los Andes.

Bradford, A. (2020). The Brussels effect: How the European Union rules the world. Oxford University Press.

Cançado Trindade, A. A. (2018). El derecho internacional de los derechos humanos en el siglo XXI (2.ª ed.). Jurídica de Chile.

Celeste, E. (2022). Digital constitutionalism: The role of Internet Bill of Rights. Routledge.

Cotino Hueso, L. (2017). Libertad de expresión e información en Internet y las redes sociales: Ejercicio, amenazas y garantías. Revista de Derecho Político, (100), 197-234. doi.org

Cotino Hueso, L. (2019). Libertad de expresión e información en internet y las redes sociales: Amenazas y riesgos en el entorno digital. Editorial Jurídica Sepín.

Cotino Hueso, L. (2023). La responsabilidad de los usuarios y plataformas por contenidos de terceros: De Delfi a Sánchez. Un endurecimiento peligroso para la libertad de expresión. Revista de Derecho Digital e Innovación.

Echeveste, M. (2015). Libertad de expresión: La doble dimensión en la jurisprudencia de la Corte Interamericana de Derechos Humanos. Editorial Juris.

Echeveste, M. (2020). Más que palabras: Libertad de expresión y discurso de odio en el Sistema Interamericano. Eunomía. Revista en Cultura de la Legalidad, (18), 320-338. doi.org

Ferrajoli, L. (2011). Principia iuris: Teoría del derecho y de la democracia (Vol. 1). Trotta.

Fiss, O. M. (2004). La ironía de la libertad de expresión. Gedisa.

Garay, A. (2016). La libertad de expresión y el discurso de odio en internet. Revista del Instituto Interamericano de Derechos Humanos, (68), 115-142.

Garay, A. (2021a). La libertad de expresión en el entorno digital: Test tripartito y nuevos desafíos. Editorial Astrea.

Garay, A. (2021b). Libertad de expresión, plataformas digitales y el desafío de la moderación privada. En Derecho, Tecnología y Sociedad (pp. 45-68). Abeledo Perrot.

Garay, A. F. (2014). La libertad de expresión en la jurisprudencia de la Corte Suprema [Tesis de maestría, Universidad de San Andrés]. Repositorio San Andrés. hdl.handle.net

García Ramírez, S. (2006). La jurisprudencia de la Corte Interamericana de Derechos Humanos. UNAM / Corte IDH.

García Roca, J. (2010). El margen de apreciación nacional en la interpretación del Convenio Europeo de Derechos Humanos. Civitas.

Grimm, D. (2009). Constitutionalism: Past, present, and future. Oxford University Press.

Habermas, J. (2010). Facticidad y validez: Sobre el derecho y el Estado democrático de derecho en términos de teoría del discurso (4.ª ed.). Trotta.

Kaye, D. (2019). Speech police: The global struggle to govern the internet. Columbia Global Reports.

Lanza, E. (2017). Estándares para una Internet libre, abierta e incluyente. Relatoría Especial para la Libertad de Expresión de la CIDH.

Lessig, L. (2006). Code: And other laws of cyberspace, version 2.0. Basic Books.

Llansó, E. (2015). Los riesgos de la delegación de la censura: El test tripartito y la moderación de contenidos privados. Estudios de Derecho Público.

Llansó, J. (2020a). Moderación de contenidos con inteligencia artificial: ¿Un nuevo paradigma para la libertad de expresión? En Derecho, Tecnología y Sociedad.

Llansó, J. (2020b). La inteligencia artificial y la moderación de los discursos de odio en internet. Sur - Revista Internacional de Derechos Humanos, 17(30), 155-163.

Loreti, D., y Lozano, G. (2014). El derecho a la comunicación: La libertad de expresión en el siglo XXI. Siglo XXI Editores.

Presno Linera, M. Á. (2014). Los derechos fundamentales como reglas. Marcial Pons.
Presno Linera, M. Á. (2017). Libertad de expresión y discursos de odio en la jurisprudencia del Tribunal Europeo de Derechos Humanos. Anuario de la Facultad de Derecho de la Universidad de Alcalá, 161-186.
Presno Linera, M. Á. (2019a). Libertades de expresión e información: Su ejercicio y sus límites. En Lecciones de Derecho Constitucional. Ediciones de la Universidad de Oviedo.
Presno Linera, M. Á. (2019b). La libertad de expresión en la jurisprudencia del Tribunal Europeo de Derechos Humanos y del Tribunal Constitucional (Segunda parte) (Cuadernos digitales de formación n.º 23). Consejo General del Poder Judicial.
Presno Linera, M. Á. (2022). La constitución digital: Derechos fundamentales y libertades públicas en la red. Editorial Marcial Pons.
Prieto Sanchís, L. (2003). Justicia constitucional y derechos fundamentales. Trotta.
Schauer, F. (2004). Las reglas en juego: Un examen filosófico de la toma de decisiones basada en reglas en el derecho y en la vida cotidiana. Marcial Pons.
Smet, S. (2017). Resolving conflicts between human rights: The case law of the European Court of Human Rights. Oxford University Press.
Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.
Sunstein, C. R. (2021). Liars: Falsehoods and free speech in an age of deception. Oxford University Press.
Tushnet, R. (2014). Power without responsibility: Intermediaries and the First Amendment. George Washington Law Review.
Voorhoof, D. (2014). The right to freedom of expression and information under the European human rights system. Robert Schuman Centre for Advanced Studies.
Waldron, J. (2005). Derecho y desacuerdos (J. L. Martí y Á. Quiroga, Trads.). Marcial Pons.
Waldron, J. (2012). The harm in hate speech. Harvard University Press.
Zagrebelsky, G. (2011). El derecho dúctil: Ley, derechos, justicia (M. Gascón, Trad.; 10.ª ed.). Trotta.

Jurisprudencia

Corte Interamericana de Derechos Humanos (Corte IDH)
Corte IDH. (1985). La Colegiación Obligatoria de Periodistas (Arts. 13 y 29 Convención Americana sobre Derechos Humanos). Opinión Consultiva OC-5/85. Serie A No. 5.
Corte IDH. (2001). Caso "La Última Tentación de Cristo" (Olmedo Bustos y otros) Vs. Chile. Sentencia de 5 de febrero de 2001 (Fondo, Reparaciones y Costas).
Corte IDH. (2006). Caso Claude Reyes y otros Vs. Chile. Sentencia de 19 de septiembre de 2006 (Fondo, Reparaciones y Costas).
Corte IDH. (2009). Caso Ríos y otros Vs. Venezuela. Sentencia de 28 de enero de 2009 (Excepciones Preliminares, Fondo, Reparaciones y Costas).
Corte IDH. (2013). Caso Mémoli Vs. Argentina. Sentencia de 22 de agosto de 2013 (Excepciones Preliminares, Fondo, Reparaciones y Costas).
Corte IDH. (2019). Caso Álvarez Ramos Vs. Venezuela. Sentencia de 30 de agosto de 2019 (Excepción Preliminar, Fondo, Reparaciones y Costas).
Tribunal Europeo de Derechos Humanos (TEDH)
TEDH. (1976). Caso Handyside v. The United Kingdom. Demanda No. 5493/72. Serie A No. 24.
TEDH. (2015). Caso Delfi AS c. Estonia [Gran Sala]. Demanda No. 64569/09. Sentencia de 16 de junio de 2015.
TEDH. (2016). Caso Magyar Tartalomszolgáltatók Egyesülete (MTE) e Index.hu Zrt c. Hungría. Demanda No. 22947/13. Sentencia de 2 de febrero de 2016.
TEDH. (2020). Caso Beizaras y Levickas c. Lituania. Demanda No. 41288/15. Sentencia de 14 de enero de 2020.

Instrumentos internacionales y otros documentos

Comité de Derechos Humanos. (2000, 18 de octubre). Malcolm Ross c. Canadá (Comunicación n.º 736/1997). ONU Doc. CCPR/C/70/D/736/1997. juris.ohchr.org
Comité de Derechos Humanos. (2011). Observación general n.º 34: Artículo 19 (Libertad de opinión y libertad de expresión) (CCPR/C/GC/34). Naciones Unidas. https://www.ohchr.org/es/treaty-bodies/ccpr/general-comments
Comité de Ministros del Consejo de Europa. (2020, 8 de abril). Recomendación CM/Rec(2020)1 sobre los impactos de los sistemas algorítmicos en los derechos humanos. Consejo de Europa. search.coe.int
Consejo de Europa. (1950, 4 de noviembre). Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales. Roma.
Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. (2012, 5 de octubre). Plan de Acción de Rabat sobre la prohibición de la apología del odio nacional, racial o religioso que constituye incitación a la discriminación, la hostilidad o la violencia. Naciones Unidas. https://www.ohchr.org/sites/default/files/Rabat_threshold_test_Spanish.pdf
Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. (2020, 15 de mayo). Prueba de umbral sobre discurso de odio. https://www.ohchr.org/es/stories/2020/05/threshold-test-hate-speech-now-available-32-languages
Organización de los Estados Americanos. (2017). Estándares para una Internet libre, abierta e incluyente. Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos. www.oas.org
Parlamento Europeo. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas sobre Inteligencia Artificial (Ley de Inteligencia Artificial). eur-lex.europa.eu
Relatoría Especial para la Libertad de Expresión (RELE). (2023a). Libertad de expresión e inteligencia artificial (OEA/Ser.L/V/II). Comisión Interamericana de Derechos Humanos. https://www.oas.org/es/cidh/expresion/informes/vigilanciaRELE.pdf
Relatoría Especial para la Libertad de Expresión (RELE). (2023b). Estándares interamericanos sobre libertad de expresión e inteligencia artificial: Guía para la moderación de contenidos en entornos digitales. Comisión Interamericana de Derechos Humanos.
Tribunal Europeo de Derechos Humanos. (2019). Guía sobre el Artículo 10 de la Convención Europea de Derechos Humanos: Libertad de expresión. Consejo de Europa. www.echr.coe.int.