



GOBERNANZA DE CONTENIDOS EN REDES SOCIALES: APROXIMACIÓN A LA INSTITUCIONALIDAD CHILENA

Content governance on social media: An approach to Chilean institutional framework

Governança de conteúdos em redes sociais: Uma aproximação à institucionalidade chilena

Javier Gallegos Gambino

Universidad de los Andes

Las Condes, Chile

jgallegos@miuandes.cl

<https://orcid.org/0009-0003-5257-8031>



<https://doi.org/10.35588/4nwxbf16>

→ Recibido

09 de enero de 2026.

→ Aceptado

11 de junio de 2026.

→ Publicado

31 de julio de 2026.

→ Cómo citar

Gallegos Gambino, J. (2026). Gobernanza de contenidos en redes sociales: Aproximación a la institucionalidad chilena. *Re-presentaciones. Periodismo, comunicación y sociedad*, (24), 24-40. <https://doi.org/10.35588/4nwxbf16>

[RESUMEN]

Este trabajo analiza el debate actual sobre la regulación de contenidos en redes sociales en línea en Chile. A partir de una revisión crítica de la literatura y de los principales modelos de moderación, se identifican, a nivel comparado, límites estructurales asociados a legitimidad democrática, transparencia y protección de derechos fundamentales. Como alternativa, se propone la gobernanza institucional de contenidos, entendida como mecanismos públicos ex post, legítimos y proporcionales. El estudio examina la experiencia chilena, destacando el rol del Consejo Nacional de Televisión y otros órganos, como un modelo con proyección al entorno digital.

[PALABRAS CLAVE]

Moderación de contenidos; libertad de expresión; redes sociales en línea; gobernanza de internet; derechos fundamentales

[ABSTRACT]

This paper analyzes the current debate on content regulation on online social media in Chile. Based on a critical review of the literature and the main moderation models, it identifies, at a comparative level, structural limits associated with democratic legitimacy, transparency, and the protection of fundamental rights. As an alternative, it proposes the institutional governance of content, understood as public, ex post, legitimate, and proportional mechanisms. The study examines the Chilean experience, highlighting the role of the National Television Council and other bodies as a model with projection into the digital environment.

[KEYWORDS]

Content moderation; freedom of expression; online social networks; internet governance; fundamental rights

[RESUMO]

Este trabalho analisa o debate atual sobre a regulação de conteúdos nas redes sociais on-line no Chile. A partir de uma revisão crítica da literatura e dos principais modelos de moderação, identificam-se, de forma comparada, limites estruturais relacionados à legitimidade democrática, à transparência e à proteção dos direitos fundamentais. Como alternativa, propõe-se a governança institucional de conteúdos, entendida como mecanismos públicos ex post, legítimos e proporcionais. O estudo examina a experiência chilena, destacando o papel do Conselho Nacional de Televisão e de outros órgãos como um modelo com potencial de projeção para o ambiente digital.

[PALAVRAS CHAVE]

Moderação de conteúdos; liberdade de expressão; redes sociais online; governança da internet; direitos fundamentais

1.- Introducción

En Chile, el estudio sobre contenidos y límites del ejercicio de la libertad de expresión en redes sociales en línea (RSL) (Del Prete y Redon Pantoja, 2020; Flores et. al., 2022; Sanandres Campis, 2023), se ha concentrado, principalmente, en el análisis de conductas nocivas específicas, tales como los discursos de odio (García, 2025), la desinformación (Jaramillo, 2024; Valenzuela, 2024; Anguita et. al, 2023), la violencia de género digital (Matus et. al, 2018); el *doxxing* (Anguita, 2021), y el derecho al olvido (Ortiz, 2025). También, hay numerosos análisis enfocados en fenómenos como el *grooming*, *ciberbullying*, *sexting* (Arias, 2018), y las llamadas “funas”¹ (Contreras y Lovera, 2021), con el objetivo de examinar sus consecuencias a nivel social, cultural y jurídico-judicial.

Con todo, pese a su impacto social y mediático, y en un contexto generalizado por la desregulación general del entorno digital a nivel regional y mundial, persiste una brecha relevante en el estudio y comprensión sistemática del concepto de moderación de contenidos (Douek, 2022; Gillespie, 2018; Grimmelman, 2015; Klonick, 2018; Peña, 2021; Roberts, 2019). Ello resulta particularmente significativo si se considera que este concepto se ha consolidado como un eje central de análisis y discusión a nivel académico, intelectual y regulatorio a escala global, tensionando los modelos regulatorios clásicos de responsabilidad por contenidos diseñados originalmente para el contexto de medios de comunicación tradicionales (Roko, 2024).

A nivel local, en Chile y Latinoamérica, existen escasas investigaciones que aporten una perspectiva sistémica sobre este fenómeno y los problemas asociados a él, en parte debido a la predominancia de enfoques académicos centrados en las experiencias y regulaciones del norte global. Por ello, resulta necesario aterrizar el análisis a la forma en que la regulación legal de los medios de comunicación tradicionales —en particular, como se verá, en el caso de la televisión— ha construido criterios, estándares y mecanismos institucionales para enfrentar problemas vinculados a la circulación de contenidos, así como al modo en que los pronunciamientos de organismos administrativos y jurisdiccionales, aun cuando no hayan sido diseñados específicamente para las redes sociales en línea, pueden incidir en la comprensión y tratamiento de estos conflictos. Este enfoque permite contar con referencias concretas para estudiar los efectos de dichas respuestas institucionales y, al mismo tiempo, proyectar una solución de mediano y largo plazo frente a las consecuencias negativas derivadas del ejercicio del poder privado de moderación de contenidos.

Frente a este escenario, la discusión global se ha centrado en dos grandes enfoques. Por un lado, se encuentra el análisis crítico de la autorregulación de las redes sociales en línea (RSL), que ha avanzado en la definición de estándares propios sobre los límites aceptables del discurso mediante la adopción de normas comunitarias y sistemas de moderación —automatizados, humanos o híbridos—, así como la creación de instancias colegiadas especializadas, como la *Oversight Board* de Meta, orientadas a revisar decisiones y establecer criterios de aplicación. Por otro lado, persiste el debate en torno al contenido y alcance de los marcos regulatorios estatales, destacando la *Digital Services Act* (DSA) de la Unión Europea y sus posibles adaptaciones al contexto latinoamericano (González Mama y Álvarez-Ugarte, 2025).

Sin embargo, dichos enfoques presentan ciertas limitaciones: el primero, carece de mecanismos democráticos de control, falta de transparencia, opacidad e, incluso, afectaciones al debido proceso en la determinación de consecuencias específicas para los usuarios (Cusumano, Gawer y Yoffie, 2021; Wong y Floridi, 2023). El segundo corre el riesgo de derivar en censura o intervencionismo excesivo por parte del Estado (Chander, 2023; Dvoskin, 2025; Shattock, 2021).

¹ En el contexto chileno, la expresión *funas* alude a una práctica de denuncia o escarnio público dirigida contra una persona u organización. Aunque su origen se vincula con actos presenciales de denuncia frente a violaciones de derechos humanos, en el entorno digital suele referirse a publicaciones en redes sociales orientadas a exponer conductas reprochadas y movilizar rechazo social.

Este ensayo propone una tercera vía, la gobernanza institucional de contenidos, acá definida como el conjunto de mecanismos ejercidos por órganos del Estado que, dotados de legitimidad pública, inciden directa o indirectamente en el ecosistema mediático y digital. Estas instituciones actúan como intermediarias entre las audiencias– usuarios/as de redes sociales en línea– y las plataformas, promoviendo un marco regulatorio mínimo, coherente con la protección de bienes jurídicos esenciales para las sociedades democráticas y con la aplicación del Derecho Internacional de los Derechos Humanos.

El caso chileno resulta especialmente ilustrativo para analizar alternativas institucionales de gobernanza de contenidos digitales. Por una parte, cuenta con una institucionalidad específica en materia audiovisual —el Consejo Nacional de Televisión—, dotada de autonomía constitucional y encargada de velar por el correcto funcionamiento de los servicios de televisión conforme al artículo 19 N° 12 de la Constitución y a la Ley N° 18.838. Por otra, diversos organismos administrativos y jurisdiccionales han comenzado a intervenir, desde sus respectivas competencias sectoriales, frente a conductas nocivas desarrolladas en redes sociales en línea.

A partir de esta experiencia, el enfoque propuesto busca enriquecer la discusión sobre gobernanza de plataformas desde una perspectiva latinoamericana. En particular, permite examinar esquemas institucionales intermedios que superen la dicotomía entre la autorregulación privada de las plataformas y la intervención estatal directa, mediante modelos dotados de legitimidad, proporcionalidad, transparencia y control público.

En este sentido, se propone avanzar hacia una arquitectura institucional integrada, capaz de articular la relación entre usuarios, plataformas y autoridades estatales, y de ofrecer respuestas más eficaces y socialmente legítimas frente a los desafíos de la comunicación digital contemporánea.

2.- Aproximaciones conceptuales a la moderación de contenidos

La moderación de contenidos se ha erigido como una dimensión estructural del funcionamiento de las plataformas digitales de RSL. De acuerdo con Peña (2021), siguiendo a Roberts (2019) y Gillespie (2018), esta práctica cumple un rol clave en la protección de la reputación corporativa de las plataformas, pues incide en la observancia de sus reglas internas y en la preservación de la participación de sus audiencias. En esta línea, Grimmelmann (2015) propuso una clasificación taxonómica de los mecanismos de moderación, poniendo de relieve su naturaleza diversa y cambiante, que abarca intervenciones como la eliminación, ocultamiento, filtrado o promoción de contenidos, según los fines y criterios que orientan a cada comunidad digital.

Por su parte, Klonick (2017) distingue entre modalidades de moderación proactiva y reactiva, aplicadas antes o después de la publicación de los contenidos, tanto por moderadores humanos o a través de sistemas automatizados. Desde un enfoque institucional, Gillespie (2018) conceptualiza la moderación como una modalidad de “custodia de las plataformas”, en virtud de la que las empresas asumen un rol decisorio sobre la visibilidad pública de ciertos contenidos.

Roberts (2019), a su turno, ha procurado visibilizar las condiciones laborales y el trabajo precarizado que promueven las plataformas a través de la moderación humana a gran escala. Por su parte, Douek (2022) cuestiona la preeminencia de la autorregulación privada y la denominada “imagen estándar” de la moderación de contenidos, advirtiendo su tendencia a simplificar el fenómeno y reducirlo a decisiones puntuales, como la eliminación de contenidos o suspensión de cuentas. Desde una perspectiva sistémica, Douek sostiene que este enfoque individualista desconoce la complejidad del entramado institucional, normativo y técnico que configura la moderación.

Finalmente, Elswah et al. (2025) señalan que los modelos actuales de moderación resultan, en términos generales, poco eficaces, en la medida de que aplican normas de alcance global que no

consideran adecuadamente las particularidades locales ni la diversidad cultural y lingüística, con efectos especialmente gravosos en los países del sur global. Frente a ello, los autores sostienen la necesidad de avanzar hacia enfoques participativos de moderación, que permitan a los usuarios cuestionar y disputar las normas que inciden en el ejercicio de sus derechos, con el respaldo de moderadores debidamente capacitados en los contextos lingüísticos y socioculturales de cada región.

En el contexto latinoamericano, se han impulsado diversas iniciativas destinadas a problematizar, conceptualizar y ordenar analíticamente los desafíos vinculados a la moderación de contenidos y a la responsabilidad de los intermediarios digitales. Entre ellos destaca el informe del programa Diálogo Latinoamericano, Moderación de contenidos y mecanismos de autorregulación: el *Oversight Board* de Facebook y sus implicancias para América Latina (Lanza y Jackson, 2021), así como el documento Mirando al Sur: Hacia nuevos consensos regionales en materia de responsabilidad de intermediarios en Internet (Del Campo et. al, 2025), presentado en el marco de una reunión de expertos del Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) en marzo de 2019. Más recientemente, en 2023, OBSERVACOM y UNESCO impulsaron el encuentro Transparencia, moderación de contenidos y libertad de expresión: perspectivas multi-actor en América Latina, centrado en el análisis y la aplicación de las Directrices para la regulación de plataformas digitales (2023), con lo cual consolidan una agenda regional que busca articular estándares de transparencia, participación y protección de la libertad de expresión en el entorno digital.

De forma más reciente, estudios del CELE (2025) han profundizado en el análisis del Trust and Safety como campo de estudio y aplicación práctica, entendido como parte de un mecanismo de gobernanza e incentivos comerciales orientado a mejorar la experiencia de los usuarios en la economía de la atención. Dichos estudios examinan esta materia desde dos perspectivas complementarias: por una parte, su compatibilidad con los estándares del Derecho Internacional de los Derechos Humanos (Del Campo, A. y Álvarez-Ugarte, R., 2025); y, por otra, el impacto de modelos regulatorios internacionales, como la *Digital Services Act*, en el escenario latinoamericano (González Mama, M. y Álvarez-Ugarte, R., 2025).

La eventual recepción latinoamericana de la *Digital Services Act* constituye, en este sentido, una novedad regulatoria relevante, pues cuenta con una ventaja comparativa frente a otros modelos: su capacidad de incidir directamente en empresas transnacionales de Internet. Sin embargo, su adaptación al contexto regional plantea dificultades específicas, derivadas del menor desarrollo de infraestructuras comunicacionales, de la existencia de estándares propios en materia de libertad de expresión y acceso a la información, y de la imposibilidad de anticipar, desde América Latina, todos los problemas que puedan surgir de la aplicación de este marco normativo en la realidad europea.

2.1.- Modelos de moderación de contenidos en RSL

La moderación de contenidos puede entenderse como un conjunto de prácticas, técnicas y decisiones que se clasifican según quién o qué decide sobre el contenido en redes sociales en línea o plataformas digitales en general. Según Alberó (2020), la moderación de contenidos puede ejecutarse de tres formas distintas: i) automatizada; ii) manual (o no automatizada); e iii) híbrida.

2.1.1.- Moderación automatizada

Este modelo se basa en el empleo de sistemas algorítmicos destinados a detectar, categorizar y, en su caso, retirar contenidos que contravienen las normas internas de las plataformas. Si bien este enfoque aporta eficiencia operativa y escalabilidad, ha sido objeto de críticas reiteradas debido a su carácter opaco, a las limitaciones para impugnar decisiones incorrectas y a su propensión a reproducir sesgos estructurales preexistentes (Cobbe, 2020; Douek, 2021).

La moderación automatizada no se limita a la detección y eliminación de contenidos, sino que abarca un conjunto más amplio de intervenciones algorítmicas destinadas a configurar la visibilidad y circulación de la información en RSL. Como señalan Pérez y Martínez (2022), dentro de estas técnicas se incluyen: la jerarquización o priorización de contenidos, que define qué publicaciones aparecen en primer lugar en los *feeds* conforme a criterios automatizados; la recomendación personalizada, sustentada en el análisis de los patrones de comportamiento de los usuarios; y la reducción de alcance, mediante el cual se limita la difusión de una publicación sin recurrir a su eliminación.

Las RSL más relevantes en la actualidad —como Facebook, TikTok, YouTube, X (Twitter) o Instagram—, debido al amplio volumen de contenidos que en ellas se comparte y difunde, dependen cada vez más de algoritmos para moderar contenido a gran escala. Herramientas como el *hash-matching* (huellas digitales de contenido) y clasificadores de *machine learning* se emplean para identificar automáticamente infracciones en áreas como discurso de odio, terrorismo o infracción de copyright, dado el volumen masivo de publicaciones generadas por usuarios (Gorwa, et. al., 2020).

No obstante, Singhal et al. (2022), señala que, aunque la moderación de contenidos realizada por humanos presenta limitaciones en términos de tiempo y escalabilidad, los sistemas automatizados basados en aprendizaje automático enfrentan obstáculos igualmente complejos, como la escasez de datos de entrenamiento adaptados a contextos regionales o variaciones lingüísticas locales. Además, sostiene que es prácticamente inviable desarrollar modelos automatizados perfectos, ya que estos deben lidiar con el contexto comunicativo, la ambigüedad del lenguaje y las formas cambiantes de violencia simbólica, a la vez que operan en entornos conflictivos que pueden ser manipulados por agentes maliciosos.

2.1.2.- Moderación manual o no automatizada

Como señalan Badouard y Bellon (2025), desde los primeros espacios de interacción en línea, distintos actores no estatales —ajenos tanto al control gubernamental como al de las propias plataformas digitales—, han tenido un rol activo en la moderación de contenidos. En comunidades virtuales más reducidas, como los grupos de *Usenet*, los blogs personales o las listas de correo, esta tarea recaía directamente en los miembros de la comunidad, quienes asumían cotidianamente la aplicación y vigilancia de las normas internas de convivencia.

El modelo de moderación no automatizada se apoya en la labor de moderadores humanos encargados de revisar contenidos reportados por usuarios o detectados por sistemas automatizados como potencialmente dañinos, ilegales o contrarios a las normas de las plataformas. Como explican Spence y DeMarco (2025), esta modalidad permite incorporar criterios contextuales que los sistemas algorítmicos no siempre capturan adecuadamente, tales como matices lingüísticos, culturales o simbólicos relevantes para evaluar la idoneidad de un contenido. Sin embargo, su implementación masiva presenta costos laborales y desafíos éticos significativos, pues estas decisiones suelen adoptarse con gran rapidez —en ocasiones entre 10 y 30 segundos— y repetirse miles de veces durante una jornada, en un entorno en que los moderadores no solo revisan material trivial, como spam, sino también imágenes, videos o textos con altos niveles de violencia, discursos de odio o abuso sexual.

Durán Allimant (2024) sostiene que la moderación humana de contenidos digitales expresa una forma de desigualdad laboral global, en tanto este trabajo esencial para las plataformas suele tercerizarse hacia países de menores ingresos, como Filipinas, India o Malasia. Allí, trabajadores capacitados y con dominio del inglés revisan contenidos para empresas y usuarios de países más ricos, bajo condiciones generalmente precarias. Esta dinámica reproduce asimetrías económicas y tecnológicas, al trasladar hacia el sur global los costos humanos de gestionar la denominada “basura digital”.

2.1.3.- Moderación basada en la comunidad (“Community Notes” de X)

Esta forma de moderación consiste en una herramienta entregada a un grupo de usuarios designados para elaborar anotaciones en publicaciones, con el propósito de contextualizarlas o contrarrestar posibles desinformaciones. Dichas anotaciones son luego valoradas por otros participantes del programa desde distintos enfoques, y aquellas que obtienen evaluaciones favorables en cuanto a su utilidad se exhiben para toda la comunidad junto a la publicación y acompañadas de un rótulo distintivo (Flynn et al., 2025). Mohammadi y Yasseri (2025) afirman que este sistema de moderación está diseñado para capitalizar la diversidad de perspectivas de los propios usuarios de la plataforma. Para ello, se muestran publicaciones específicas a quienes participan en el programa, solicitándoles redactar notas que aporten contexto a dichas publicaciones.

En enero de 2021, la red social en línea Twitter –actualmente denominada X– puso en marcha una iniciativa piloto de moderación de contenidos conocida como *Birdwatch*, posteriormente consolidada bajo el nombre de *Community Notes*. Este mecanismo, considerado uno de los primeros intentos de moderación comunitaria a gran escala, tuvo como propósito descentralizar las decisiones de moderación mediante un sistema basado en la inteligencia colectiva de los usuarios, orientado a permitir la incorporación colaborativa de notas explicativas en publicaciones potencialmente engañosas. En su fase inicial, el programa operó de manera experimental y en un entorno aislado, con el objetivo de emular procesos de verificación colaborativa y alcanzar niveles de precisión comparables a los de la revisión realizada por expertos humanos (Saeed et al., 2022; Mohammadi et al., 2025), promoviendo así un modelo participativo de intervención sobre el contenido sin recurrir directamente a su eliminación.

De acuerdo con el *Center for Democracy & Technology* (2025), los mecanismos de recopilación de conocimiento basados en múltiples perspectivas comunitarias no constituyen una innovación radical, siendo Wikipedia un ejemplo paradigmático de su aplicación exitosa. En esta línea, estudios recientes sobre el funcionamiento de las *Community Notes* de X arrojan resultados prometedores: cuando las notas incorporan fuentes verificadas y confiables, tienden a generar mayores niveles de consenso entre usuarios de distintas orientaciones ideológicas y, adicionalmente, su presencia puede tener un efecto disuasivo respecto de la recirculación de contenidos engañosos (Kangur et al., 2024; Chuai et al., 2024).

No obstante, autores como Yasseri y Menczer (2023) han advertido que la red que conecta a autores y evaluadores de notas reproduce la fuerte polarización existente en la red de amistades en X. En la misma línea, Allen et al. (2022) evidenciaron que los usuarios tienden a escribir notas sobre publicaciones contra partidarios, calificarlas como engañosas y considerar poco útiles las notas redactadas por estos, mientras que valoran como útiles aquellas provenientes de copartidarios. Asimismo, Zhao (2025), sostiene que la creciente politización de X y de Musk provocó un éxodo de usuarios que debilitó la legitimidad de la moderación colaborativa: aunque influyente en la esfera digital, la plataforma se ha convertido en un espacio de desinformación y odio, lo cual evidencia que la gobernanza arbitraria puede ser más dañina que beneficiosa.

3.- Hacia una gobernanza institucional de contenidos

En la discusión actual sobre la regulación de los contenidos en RSL, las posiciones dominantes suelen ubicarse entre dos polos opuestos: de una parte, la autorregulación empresarial, ejercida por plataformas privadas a través de normas comunitarias, sistemas algorítmicos y decisiones internas; y, de otra, la regulación estatal que establece límites jurídicos mediante leyes, sanciones o mecanismos judiciales. Ambas alternativas han evidenciado déficits relevantes: la primera, debido a su opacidad, discrecionalidad y escasos niveles de rendición de cuentas; la segunda, por riesgos de injerencia estatal indebida o restricciones desproporcionadas a la libertad de expresión. Este último

aspecto es examinado por Alkiviadou (2025), quien advierte que estos modelos tienden a privatizar decisiones sobre la legalidad de los contenidos sin garantías suficientes, incentivando la sobremoderación y erosionando los principios de necesidad y proporcionalidad propios del Derecho Internacional de los Derechos Fundamentales.

Ante este panorama, se propone la noción de gobernanza institucional de contenidos, como una alternativa intermedia o tercera vía, inserta en el ámbito de la acción estatal, pero diferenciado tanto de la regulación legislativa estricta como de la autorregulación excesiva de las plataformas privadas. Se trata del conjunto de estrategias, prácticas y mecanismos desarrollados por instituciones públicas para incidir, orientar, observar o intervenir de manera legítima en la circulación de contenidos digitales, con el fin de garantizar derechos fundamentales, proteger bienes jurídicos colectivos y fortalecer el ecosistema informativo.

A diferencia de la “moderación de contenidos”, término habitualmente asociado a la acción privada de las plataformas (como eliminar o clasificar contenidos según rankings algorítmicos), la gobernanza institucional no busca operar sobre cada publicación individual ni necesariamente imponer sanciones directas. Su función es más amplia: abarca la producción de lineamientos normativos, acciones de fiscalización, generación de estándares ético-comunicacionales, procedimientos de mediación entre usuarios y plataformas, e incluso la producción de evidencia sobre los efectos sociales de determinados contenidos y el impacto de estos mecanismos en la disuasión de conductas nocivas.

Este tipo de gobernanza no implica promover un control estatal de la producción de contenidos en RSL, sino más bien una forma de intervención democrática, legítima y proporcional, en el marco de principios como el interés público, la transparencia, la legalidad y la participación; coherentes también con la aplicación de los criterios de imperiosidad, necesidad e idoneidad del test tripartito promovido por la Convención Americana de Derechos Humanos (González Mama, M. y Álvarez-Ugarte, R. 2025).

El caso chileno resulta ilustrativo para probar esta hipótesis. Sin tener una ley general sobre redes sociales en línea ni plataformas digitales, el país ya cuenta con una serie de instituciones públicas que ejercen funciones de gobernanza en relación con contenidos emanados de medios de comunicación masivos, en particular, de la televisión), e incluso de contenidos creados y difundidos en RSL, aunque de manera atomizada e inorgánica. El desafío de este modelo supone avanzar hacia una arquitectura institucional articulada, que opere como puente entre los usuarios y las plataformas y sea respetuosa de los estándares internacionales de derechos humanos.

3.1.- La experiencia chilena

A lo largo de las últimas décadas, el Estado chileno ha sido condenado por la Corte Interamericana de Derechos Humanos en múltiples ocasiones por vulneraciones al derecho a la libertad de expresión. Casos emblemáticos contra Chile ante la Corte Interamericana de Derechos Humanos permiten identificar distintos estándares sobre libertad de expresión. En *La Última Tentación de Cristo* (2001), el tribunal condenó la censura previa derivada de la prohibición de exhibir una obra cinematográfica; en *Palamara Iribarne* (2005), cuestionó tanto la prohibición de publicar un libro crítico como el uso de sanciones penales y militares frente a su difusión; y en *Claude Reyes* (2006), reconoció el acceso a la información pública como una dimensión inherente al derecho a buscar y recibir información.

Estos antecedentes demuestran que cualquier modelo de gobernanza institucional de contenidos adoptado por Chile debe ajustarse estrictamente a los estándares del Sistema Interamericano. Según la Convención, la censura previa está prohibida salvo excepciones, y solo se permiten responsabilidades ulteriores si están establecidas por ley, persiguen fines legítimos y son necesarias

y proporcionales. Incluso medidas adoptadas con fines aparentemente protectores —como sancionar desinformación, proteger la honra o garantizar el pluralismo— deben evitar derivar en censura encubierta o efectos inhibitorios desproporcionados sobre el debate público, pues esto podría comprometer nuevamente la responsabilidad internacional del Estado.

En ese marco, la institucionalidad chilena ofrece una alternativa para observar cómo la gobernanza de contenidos puede materializarse sin necesidad de una legislación específica sobre RSL. A través de diferentes organismos públicos, el Estado ha ido desplegando progresivamente capacidades regulatorias, orientadoras y fiscalizadoras en relación a los contenidos digitales.

3.1.1.- El rol del Consejo Nacional de Televisión (CNTV)

El CNTV constituye un órgano colegiado que goza de autonomía constitucional, cuya legitimidad y eficacia lo posicionan como un referente institucional en Chile. Su labor se funda en un marco normativo sólido y se ejerce conforme a los principios de legalidad, proporcionalidad y debido proceso, permitiéndole aplicar controles ex post con respaldo judicial.

Su habilitación está reconocida expresamente por la Constitución Política de Chile y la Ley N°18.838 para moderar contenidos televisivos luego de que éstos han sido emitidos (facultades de fiscalización y supervisión), con el objetivo de iniciar el procedimiento administrativo sancionatorio que busque determinar eventuales responsabilidades de los servicios televisivos, con miras al resguardo institucional del ámbito colectivo de la libertad de expresión, y el permanente respeto a los bienes jurídicos tutelados y descritos en el artículo 1° de la Ley N°18.838.

El Tribunal Constitucional de Chile, en diversos pronunciamientos, ha reconocido la relevancia del rol que desempeña el CNTV en la moderación de contenidos televisivos, así como la conformidad de su marco regulatorio y potestades fiscalizadoras con la Constitución. Esta validación no sólo se expresa en el reconocimiento expreso de que las atribuciones del CNTV se orientan a prevenir la arbitrariedad en la aplicación de sanciones a servicios de televisión, sino también en la valoración institucional que se otorga a la televisión como medio que, por su masividad y potencial formativo, requiere de un régimen jurídico especial y mecanismos de control que resguarden el correcto funcionamiento del sistema comunicacional.

Así, destaca, en primer lugar, que la televisión, a diferencia de los otros medios de comunicación social, está subordinada a un correcto funcionamiento, definido por el legislador. Esto es un conjunto de estándares que la televisión no contradecir o no promover, bajo sanción del Consejo Nacional de Televisión. En segundo lugar, que el CNTV es un organismo autónomo, integrado por un colectivo de integración pluralista, con facultades normativas, de fiscalización y de sanción. Y, en tercer lugar, que la televisión está sujeta a una serie de cargas y obligaciones que no tienen los otros medios, debido a su poder de penetración y masividad.

Aunque el CNTV no interviene directamente en el ámbito de las redes sociales en línea (RSL), su modelo de fiscalización ex post, su autonomía constitucional y su enfoque garantista lo posicionan como un referente plausible de gobernanza institucional aplicable al entorno digital. Su capacidad para recibir denuncias ciudadanas, emitir criterios interpretativos y aplicar sanciones con efecto disuasivo configura un marco de responsabilidad comunicacional que podría proyectarse —con las adaptaciones correspondientes— al ecosistema de plataformas digitales. Esto se ve reforzado por el carácter eminentemente técnico del organismo, cuya estructura orgánica contempla la existencia de departamentos con profesionales calificados que efectúan un análisis riguroso de los antecedentes y la deliberación posterior de un consejo colegiado compuesto por 11 miembros —nombrados por el Senado, según las reglas que establece la ley— cuyas resoluciones, en todo caso, son susceptibles de revisión judicial posterior a través de la interposición de recursos de ilegalidad administrativa ante tribunales superiores de justicia.

3.1.2.- Otras instituciones relevantes

Si bien el CNTV constituye el principal referente institucional en Chile en materia de regulación de contenidos audiovisuales, no es el único organismo que ha intervenido frente a los efectos nocivos de ciertos discursos difundidos en entornos digitales. Diversas instituciones públicas, desde sus competencias sectoriales, han comenzado a incidir en el ecosistema informativo digital mediante labores de fiscalización, interpretación normativa, advertencias públicas y, en algunos casos, potestades sancionatorias.

Estas intervenciones se advierten, por ejemplo, en materias de protección del consumidor, publicidad digital, información financiera, salud pública y tutela judicial de derechos fundamentales. En estos ámbitos, organismos como el SERNAC², la CMF³, las autoridades sanitarias y los tribunales de justicia⁴ han desarrollado criterios orientados a enfrentar prácticas engañosas, contenidos potencialmente perjudiciales, asesorías no autorizadas, discursos lesivos para la honra o la vida privada, y otras formas de afectación derivadas de la circulación de contenidos en redes sociales en línea.

Aunque estas actuaciones no configuran propiamente un sistema de moderación de contenidos, sí revelan una progresiva sensibilidad institucional frente a los riesgos asociados al ecosistema digital. En particular, muestran que la circulación de discursos en plataformas digitales puede comprometer bienes jurídicos relevantes, como la salud, la información veraz, la protección de los consumidores, la honra, la imagen y la vida privada.

Con todo, el problema central radica en que estas respuestas se producen de manera fragmentada, reactiva y sectorial. Al no existir una política común ni mecanismos estables de coordinación interinstitucional, las intervenciones públicas tienden a operar caso a caso, sin conformar todavía un modelo coherente de gobernanza institucional de contenidos digitales. Esta dispersión confirma la necesidad de avanzar hacia marcos regulatorios e institucionales capaces de abordar la complejidad del fenómeno sin reducirlo a respuestas aisladas ni a formas desproporcionadas de intervención estatal.

4.- Conclusiones

A lo largo de este ensayo hemos sostenido que el Estado se encuentra en condiciones de ofrecer respuestas legítimas y eficaces frente a las externalidades negativas que caracterizan al ecosistema digital, particularmente en lo relativo a la circulación y regulación de contenidos en RSL. Aunque las plataformas desempeñan un rol central en las tareas de moderación, la envergadura y complejidad de los desafíos actuales hacen necesario avanzar hacia formas complementarias de intervención pública que aseguren una protección efectiva de los derechos fundamentales.

En este contexto, se sugiere el desarrollo del concepto de “gobernanza institucional de contenidos”, como una vía intermedia entre la autorregulación privada y la regulación estrictamente legislativa, en la cual el Estado actúa mediante instituciones técnicas, colegiadas, fiscalizables y sometidas a control judicial. El examen de la experiencia del Consejo Nacional de Televisión (CNTV) chileno,

² En 2018, a través de un estudio de consumidores, se definió la categoría de *influencer* como un actor con capacidad de incidir en decisiones de consumo, destacando su rol como prescriptor ante audiencias segmentadas (SERNAC, 2018). Esta visión fue luego profundizada en 2022 mediante una circular interpretativa sobre publicidad nativa e *influencers*, que amplió su definición y agregó ciertos estándares para transparentar contenidos patrocinados.

³ En agosto de 2023 (CMF, 2023), emitió una advertencia pública sobre los riesgos de los llamados “*influencers* financieros”, caracterizados como aquellos que inducen a las audiencias a inversiones de alto riesgo sin respaldo legal ni información veraz. Además de su función fiscalizadora, la CMF cumple un rol preventivo y educativo, y así se configura como un actor relevante en el ámbito de la gobernanza.

⁴ La Corte Suprema, en diversos fallos (Roles N°144.290-2020, N°90.737-2020, N°157.732-2022 y N°179.399-2022), ha establecido que el ejercicio de la libertad de expresión encuentra límites en la protección de la honra, la vida privada y la imagen personal, especialmente cuando expresiones difundidas en redes sociales atribuyen conductas ilícitas o deshonrosas, sin respaldo judicial.

permite identificar un modelo de intervención estatal que, sin recurrir a la censura previa, ha logrado establecer estándares de responsabilidad comunicacional en medios tradicionales. Ello se ha realizado a partir de una lógica correctiva, respetuosa del debido proceso, basada en criterios de proporcionalidad y dotada de mecanismos institucionales de rendición de cuentas. Este diseño resulta proyectable al entorno digital, especialmente en materias vinculadas a la regulación de contenidos nocivos, la transparencia en procesos de moderación y la protección de bienes jurídicos particularmente sensibles en democracias contemporáneas, como la libertad de expresión, el pluralismo o los derechos de niños, niñas y adolescentes.

Asimismo, el trabajo ha desagregado el concepto de moderación de contenidos en sus principales variantes (humana, automatizada e híbrida), destacando las fortalezas y limitaciones propias de cada una. La moderación humana aporta contexto y mayores garantías, pero enfrenta serios problemas de escalabilidad; la automatizada ofrece eficiencia operativa, pero presenta déficits relevantes de transparencia y riesgos de sesgos algorítmicos; y la modalidad híbrida emerge como una solución intermedia crecientemente adoptada por las plataformas, cuya eficacia de todas formas está mediada por la operatividad de algoritmos y normas internas de las mismas. Con todo, el problema transversal radica en que estas formas de moderación operan, en la actualidad, sin una supervisión externa robusta, lo que se traduce en déficits de legitimidad y en mecanismos de apelación insuficientes para los usuarios.

En este escenario, la gobernanza institucional de contenidos, entendida como la intervención activa del Estado y en el diseño de criterios, supervisión y evaluación de sistemas de moderación, se presenta como una alternativa que no busca afectar la autonomía de las plataformas, sino establecer estándares mínimos comunes orientados a la protección de los Derechos Humanos.

Finalmente, si bien Chile ha sido objeto de cuestionamientos en el ámbito interamericano por eventuales vulneraciones a la libertad de expresión, también ha desarrollado jurisprudencia que reconoce la necesidad de armonizar este derecho con otros bienes constitucionales. El desafío, por tanto, no consiste en abdicar del rol estatal, sino en perfeccionar su ejercicio para hacerlo más transparente, técnico y garantista. En este sentido, la propuesta de gobernanza institucional de contenidos puede entenderse también en un plano operacional: a través de la creación de una institucionalidad especializada dedicada a la supervisión y evaluación de los sistemas de moderación, o a través de la articulación y coordinación de las agencias públicas que ya han dado pasos firmes en la protección de derechos de las personas en el entorno digital. El objetivo, entonces, consiste en dotar al Estado de capacidades que permitan establecer estándares comunes, ofrecer vías efectivas de rendición de cuentas y fortalecer la autonomía operativa de las plataformas sin comprometer ni afectar la libertad de expresión.

Futuras investigaciones podrán incorporar evaluaciones empíricas sobre los efectos de los distintos modelos de gobernanza institucional, en relación con su alcance y eventual tensión con sistemas de moderación privados de contenido. Ello supone evaluar comparativamente las experiencias regulatorias, sus decisiones y percepciones de legitimidad de usuarios y audiencias frente al establecimiento de criterios.

5.- Referencias bibliográficas

- Albero, J. (2020). *La libertad de expresión ante la sociedad algorítmica: una mirada al problema de la moderación de contenidos online*. Documento de trabajo de la Cátedra José María Cervelló, AJ8-259. IE University. <https://catedracervello.ie.edu/wp-content/uploads/sites/570/2021/02/AJ8-259.pdf>
- Alkiviadou, N. (2025). *Moderating hate or moderating rights? The paradox of the European approach to online hate speech and platform liability*. SSRN Electronic Journal <https://doi.org/10.2139/ssrn.5432875>
- Allen, J., Martel, C., y Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article 341 1–19. <https://doi.org/10.1145/3491102.3502040>
- Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P., y Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter: Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info*, (49), 98–124. <https://doi.org/10.7764/cdi.49.27817>
- Anguita, P. (2021). Freedom of Expression in Social Networks and Doxing. En L. Corredoira, I. B. Mallén y R. C. Presuel (Eds.) *The Handbook of Communication Rights, Law, and Ethics* (pp. 279-291). John Wiley & Sons. <https://doi.org/10.1002/9781119719564.ch23>
- Arias Cerón, M., Buendía, L., y Fernández Palomares, F. (2018). Grooming, cyberbullying y sexting en estudiantes en Chile según sexo y tipo de administración escolar. *Revista chilena de pediatría*, 89(3), 352-360. <https://doi.org/10.4067/S0370-41062018005000201>
- Badouard, R., y Bellon, A. (2025). Introduction to the special issue on content moderation on digital platforms. *Internet Policy Review*, 14(1), 1-24. <https://policyreview.info/articles/analysis/introduction-content-moderation-digital-platforms>
- Caso “La Última Tentación de Cristo” (Olmedo Bustos y otros) vs. Chile. Fondo, reparaciones y costas (2001, 5 de febrero). Corte Interamericana de Derechos Humanos (Carlos Vicente de Roux Rengifo, Juez). https://www.corteidh.or.cr/docs/casos/articulos/seriec_73_esp.pdf
- Caso Claude Reyes y otros vs. Chile. Fondo, reparaciones y costas. (2006, 19 de septiembre). Corte Interamericana de Derechos Humanos (Sergio García Ramírez, Juez). https://www.corteidh.or.cr/docs/casos/articulos/seriec_151_esp.pdf
- Caso Palamara Iribarne vs. Chile. Fondo, reparaciones y costas. (2005, 22 de noviembre). Corte Interamericana de Derechos Humanos (Antônio Augusto Cançado Trindade, Juez). https://www.corteidh.or.cr/docs/casos/articulos/seriec_135_esp.pdf
- Caso Urrutia Laubreaux vs. Chile. Fondo, reparaciones y costas. (2020, 27 de agosto). Corte Interamericana de Derechos Humanos (Eugenio Raúl Zaffaroni, Juez). https://www.corteidh.or.cr/docs/casos/articulos/seriec_409_esp.pdf
- Center for Democracy and Technology. (2025). *Making Meta's Community Notes work: Current challenges and opportunities*. <https://cdt.org/insights/making-metas-community-notes-work-current-challenges-and-opportunities/>
- Chander, A. (2023). When the Digital Services Act Goes Global. *Berkeley Technology Law Review*, 28(3), 1067-1088. <https://scholarship.law.georgetown.edu/facpub/2548>

- Chuai, Y., Pilarski, M., Lenzini, G., y Pröllochs, N. (2024). *Community notes reduce the spread of misleading posts on X*. *OSF Preprints*. <https://doi.org/10.31219/osf.io/3a4fe>
- Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 34, 739–766. <https://doi.org/10.1007/s13347-020-00429-0>
- Comisión para el Mercado Financiero. (2023). *CMF alerta sobre riesgos de seguir recomendaciones de “influencers financieros” en redes sociales*. <https://www.cmfchile.cl/portal/prensa/615/w3-article-82812.html>
- Comité Asesor contra la Desinformación. Ministerio de Ciencia, Tecnología, Conocimiento e Innovación. (2023). *El fenómeno de la desinformación: Revisión de experiencias internacionales y en Chile*. <https://drive.google.com/file/d/1Onkr25GT1lBux7Uvblmc6-kkNclAg6Aj/view>
- Contreras Vásquez, P., y Lovera Parmo, D. (2021). Redes sociales, funas, honor y libertad de expresión: análisis crítico de los estándares de la jurisprudencia de la Corte Suprema chilena. *Derecho PUCP*, (87), 345–371. <https://doi.org/10.18800/derechopucp.202102.010>
- Cusumano, M. A., Gawer, A., y Yoffie, D. B. (2021). *Can self-regulation save digital platforms?* *Questrom School of Business, Boston University*. https://questromworld.bu.edu/platformstrategy/wp-content/uploads/sites/49/2021/07/PlatStrat2021_paper_38.pdf
- Del Campo, A., y Álvarez-Ugarte, R. (2025). *Trust and Safety's Blindspots: A Latin American Perspective (CELE Research Paper No. 68)*. Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE). <https://ssrn.com/abstract=5355934>
- Del Campo, A., Zara, N., y Álvarez-Ugarte, R. (2025). Are risks the new rights? The perils of risk-based approaches to speech regulation. *Journal of Intellectual Property, Information Technology and Electronic Commerce*, 16(2), 1-15. <https://doi.org/10.2139/ssrn.5161173>
- Del Prete, A., y Redon Pantoja, S. (2020). Las redes sociales on-line: Espacios de socialización y definición de identidad. *Psicoperspectivas*, 19(1), 1–11. <https://doi.org/10.5027/psicoperspectivas-vol19-issue1-fulltext-1834>
- Douek, E. (2020). Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *Colombia Law Review*, 121(3), 1-76. <https://ssrn.com/abstract=3679607>
- Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, 136(2), 526–607. <https://www.jstor.org/stable/27339278>
- Durán Allimant, R. (2024). Trabajo digital y “sonambulismo tecnológico”: el caso de la moderación de contenido en Internet. *Athenea Digital. Revista De Pensamiento E Investigación Social*, 24(1), 1-18. <https://doi.org/10.5565/rev/athenea.3376>
- Dvoskin, B. (2024). The Illusion of Inclusion: The False Promise of the New Governance Project for Content Moderation. *Fordham Law Review*, 93, 1-61. <https://ssrn.com/abstract=4785833>
- Elsawah, M., Bhatia, A., y Thakur, D. (2025). *Content moderation in the Global South: A comparative study of four low-resource languages*. Center for Democracy and Technology. <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/>
- Flores Flores, A. J., Bernal González, I., y Sevilla Salazar, V. (2022). Redes sociales y éxito comercial: diseño de un modelo teórico bajo el contexto de la pandemia global COVID-19. *Vinculatégica EFAN*, 8(1), 42–52. <https://doi.org/10.29105/vtga8.1-230>

- Flynn, A., Vakhitova, Z., Wheildon, L., Harris, B., y Robards, B. (2025). Content Moderation and Community Standards: The Disconnect Between Policy and User Experiences Reporting Harmful and Offensive Content on Social Media. *Policy & Internet*, 17, 1-13. <https://doi.org/10.1002/poi3.70006>
- García-García, J. (2025). Respuestas normativas para enfrentar el discurso del odio desde la experiencia europea. *European Public & Social Innovation Review*, 10, 1–13. <https://doi.org/10.31637/epsir-2025-1904>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- González Mama, M. (2024). *A propósito de la metáfora de las redes sociales como foro público en los Estados Unidos (Reporte Técnico No. 60)*. Centro de Estudios en Libertad de Expresión (CELE). https://www.palermo.edu/Archivos_content/2024/cele/julio/paper-foro-publico/240710-a-proposito-metafora-de-rss.pdf
- González Mama, M., y Álvarez-Ugarte, R. (2025). *Regulatory modeling: Tracking the influence of the Digital Services Act in Latin America (Research Paper No. 63)*. Centro de Estudios en Libertad de Expresión (CELE). https://www.palermo.edu/Archivos_content/2025/cele/abril/2025_04_24_Regulatory_Modeling.pdf
- Gorwa, R., Binns, R., y Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1). <https://doi.org/10.1177/2053951719897945>
- Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology*, 17(1), 42–83. <https://james.grimmelmann.net/files/articles/virtues-of-moderation.pdf>
- Jaramillo Castro, O. (2024). Caracterización del flujo informativo de la desinformación en redes sociales: El caso chileno en Twitter entre 2018 y 2022. En I. Serrano y L. Corredoira (Eds). *Democracia y Desinformación. Nuevas formas de polarización, discursos de odio y campañas en redes. Respuestas regulatorias de Europa y América Latina* (pp. 121 – 137). Editorial Dykinson. <https://observatoriodesinformacion.cl/wp-content/uploads/2024/03/Democracia-y-desinformacion-121-137.pdf>
- Kangur, U., Chakraborty, R., y Sharma, R. (2026). Who Checks the Checkers? Exploring Source Credibility in Twitter’s Community Notes. *Journal of Computational Social Science*, 9(24). <https://doi.org/10.1007/s42001-025-00455-y>
- Kian [naturelvr49] (Recuperado el 15 de septiembre de 2025). *Introduction. Community Notes Guide*. X. (s. f.). <https://communitynotes.x.com/guide/en/about/introduction>
- Klonick, K. (2018). The new governors: the people, rules and processes governing online speech. *Harvard Law Review*, 131(6), 1598–1670. <http://www.jstor.org/stable/44865879>
- Lanza, E., y Jackson, M. (2021). *Moderación de contenidos y mecanismos de autorregulación: El Oversight Board de Facebook y sus implicancias para América Latina*. El Diálogo Interamericano. <https://thedialogue.org/analysis/moderacion-de-contenidos-y-mecanismos-de-autorregulacion-el-oversight-board-de-facebook-y-sus-implicancias-para-america-latina?lang=es>
- López-Belmonte, J., Pozo-Sánchez, S., y Fuentes-Cabrera, A. (2020). La realidad de la aplicación de redes sociales en el entorno educativo: El caso de una cooperativa de enseñanza de Ceuta. *Revista Electrónica Educare*, 24(1), 165–185. <https://doi.org/10.15359/ree.24-1.17>

- López-Sáez, M. Á., Pérez-Torres, V., Pastor, Y., Lobato-Rincón, L. L., Thomas, H., y Angulo-Brunet, A. (2024). Social media uses amongst adolescents: motives, minority stress and eudaimonic well-being. *International Journal of Clinical and Health Psychology*, 24(2). <https://doi.org/10.6018/analesps.556871>
- Matus, J., Rayman, D., y Vargas, R. (2018). *Violencia de género en internet en Chile*. Fundación Datos Protegidos. <https://datosprotegidos.org/descarga-violencia-de-genero-en-internet-en-chile/>
- Mohammadi, S., y Yasseri, T. (2025). *AI feedback enhances community-based content moderation through engagement with counterarguments* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2507.08110>
- Ortiz Mesías, L. (2025). El derecho al olvido digital en Chile ¿un nuevo derecho fundamental?. *Estudios constitucionales*, 23(1), 129-156. <https://doi.org/10.4067/s0718-52002025000100129>
- Peña Jiménez, P. J. (2021). Entre analogías y metáforas: El debate sobre la moderación de contenidos en las redes sociales. *Revista de las Cortes Generales*, (111), 265–311. <https://doi.org/10.33426/rcg/2021/111/1614>
- Pérez Bertrán, A. L., y Martínez Elebi, C. (2022). *Moderación privada de contenidos en internet y su impacto en el periodismo*. Observatorio Latinoamericano de Regulación, Medios y Convergencia (Observacom). <https://www.observacom.org/wp-content/uploads/2022/10/Moderacion-y-Periodismo.pdf>
- Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press. <https://doi.org/10.2307/j.ctvhrcz0v>
- Roko, P. (2025). *Old standards, new challenges: Keys to addressing internet disinformation in Inter-American jurisprudence* (CELE Research Paper No. 63). Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE). <https://ssrn.com/abstract=4758336>
- Saeed, M., Traub, N., Nicolas, M., Demartini, G., y Papotti, P. (2022). Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? En M. A. Hasan y L. Xiong. (General Chairs), *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1736–1746). ACM. <https://doi.org/10.1145/3511808.3557279>
- Sanandres Campis, E. (2023). Aplicación del análisis de redes sociales para el estudio de las redes de comunicación en línea: evidencia empírica de Twitter. *Empiria: Revista de Metodología de las Ciencias Sociales*, 57, 165–188. <https://doi.org/10.5944/empiria.57.2023.36434>
- Sentí Navarro, C. (2024). Desórdenes informativos en línea y la moderación de contenidos en la era digital. *Revista de Derecho Político*, (121), 203–233. <https://doi.org/10.5944/rdp.121.2024.43073>
- Servicio Nacional del Consumidor. (2018). *Estudio de publicidad online: El caso de la publicidad nativa*. https://www.sernac.cl/portal/619/articles-53024_archivo_01.pdf
- Shattock, E. (2021). Self-regulation 2.0? A critical reflection of the European fight against disinformation. *Harvard Kennedy School Misinformation Review*, 2(4), 1–10. <https://misinforeview.hks.harvard.edu/article/self-regulation-20-a-critical-reflection-of-the-european-fight-against-disinformation/>
- Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., y Stringhini, G. (2023). SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. *IEEE*

- Spence, R., y DeMarco, J. (2025). Content moderator mental health and associations with coping styles: Replication and extension of previous studies. *Behavioral Sciences*, 15(4), 487. <https://doi.org/10.3390/bs15040487>
- Valenzuela, S. (2024). Cuando los algoritmos son editores: Cómo las redes sociales, la IA y la desinformación alteran el consumo de noticias. *Comunicación y medios*, 33(49), 186-191. <https://doi.org/10.5354/0719-1529.2024.74976>
- Wong, D., y Floridi, L. (2023). Meta's Oversight Board: A Review and Critical Assessment. *Minds & Machines*, 33, 261–284. <https://doi.org/10.1007/s11023-022-09613-x>
- Yasseri, T., y Menczer, F. (2023). Can crowdsourcing rescue the social marketplace of ideas? *Communications of the ACM*, 66(9), 42–45. <https://doi.org/10.1145/3603620>
- Zhao, A. D. (2025). *Checking, moderating, adapting: Collective engagement for digital integrity* [Doctoral dissertation, Cornell University]. ProQuest Dissertations & Theses Global (UMI No. 32169297). <https://ecommons.cornell.edu/bitstreams/d5bb2837-62a2-4e8b-a9e0-d2b8132c1f72/download>

6.- Contribución de autores

Javier Gallegos Gambino: Conceptualización; Investigación; Metodología; Validación; Redacción – borrador original; Redacción – Revisión y Edición.