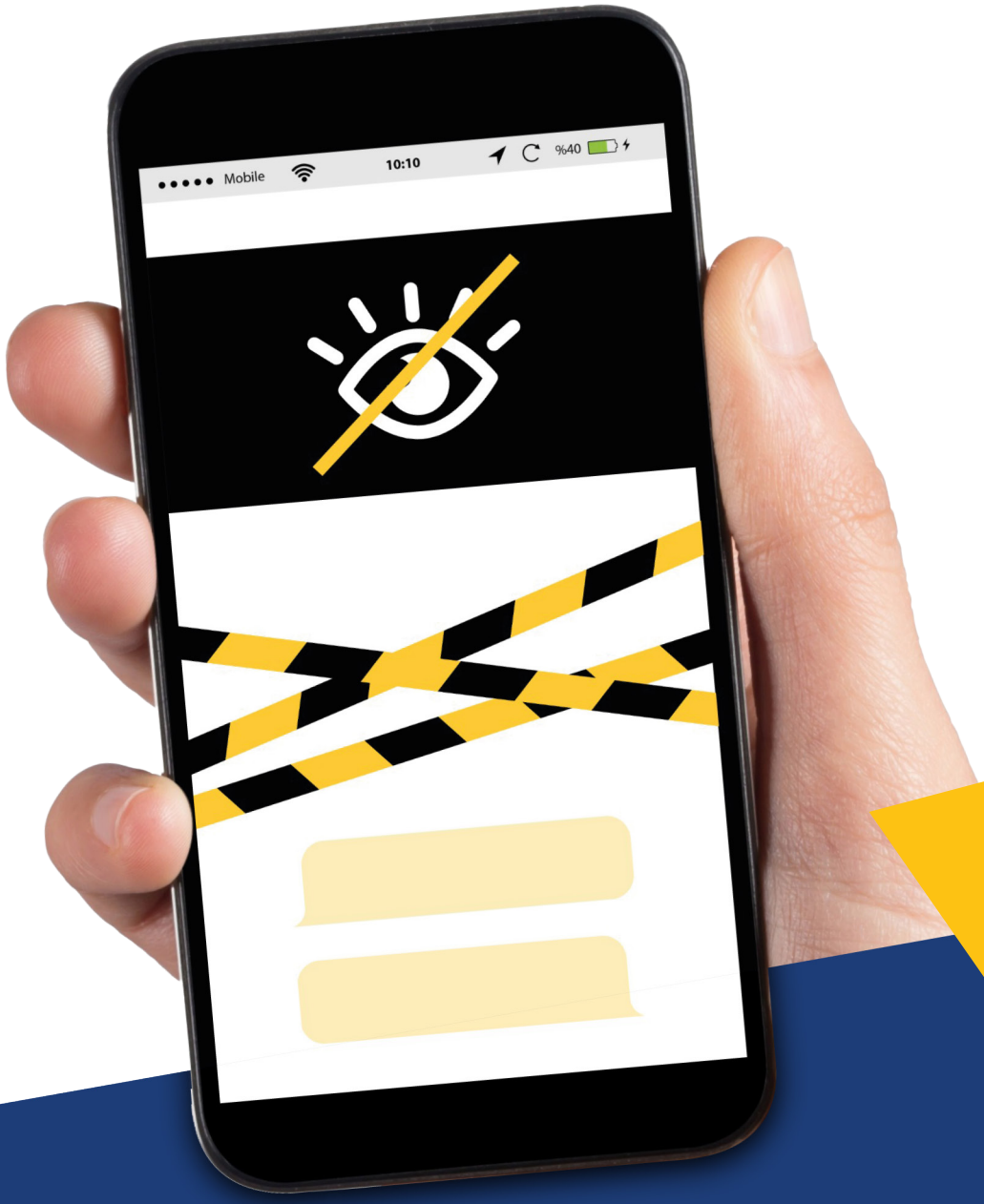

EL SHADOW BANNING:
la censura sutil y oculta de las grandes plataformas digitales

OBSERVACOM

El shadow banning: la censura sutil y oculta de las grandes plataformas digitales

Estudio sobre las prácticas de reducción de alcance de contenidos y cuentas en redes sociales y su relación con la libertad de expresión en línea

Es una publicación de
OBSERVACOM
Observatorio Latinoamericano de Regulación, Medios y Convergencia

Av. Libertador 1878 apto. 715
Montevideo, Uruguay
www.observacom.org

Con el apoyo de

y el apoyo de

Carolina Martínez Elebi - Autora

Es Licenciada en Ciencias de la Comunicación por la Universidad de Buenos Aires (UBA) de Argentina, donde se desempeña como docente desde 2011. Es consultora en temas vinculados al impacto de las tecnologías de la información y la comunicación en los derechos humanos y es directora del medio digital DHyTecno. En 2018 cursó el Programa de Derecho de Internet y Tecnologías de las Comunicaciones del Centro de Estudios en Tecnología y Sociedad (CETyS-UdeSA). Es coordinadora académica de la Diplomatura Superior en Inteligencia Artificial y Sociedad de la Universidad de Tres de Febrero (UNTREF) e integra el Observatorio de Impactos Sociales de la Inteligencia Artificial de esa misma universidad.

Vladimir Cortés Roshdestvensky - Autor

Es especialista en derechos humanos con más de una década de experiencia en derechos digitales, libertad de expresión y gobernanza democrática. Con un Máster en Derechos Humanos por la Universidad de Padua, ha liderado investigaciones sobre IA, brecha digital y moderación de contenidos. Actualmente es Director de Campañas y Alianzas en Digital Action, donde coordina estrategias para exigir rendición de cuentas a gobiernos y empresas tecnológicas. Recibió la beca de Líderes de LACNIC y «Nicola Tonon». Trabajó en ARTICLE 19 y como analista para el informe Freedom on the Net de Freedom House.

01.

INTRODUCCIÓN

En el ecosistema digital actual, las plataformas ejercen un rol central como intermediarias en la circulación de información. Como parte de esa función, implementan sistemas de moderación de contenidos que incluyen medidas visibles y relativamente conocidas, como la eliminación de publicaciones, la suspensión temporal o definitiva de cuentas y otras sanciones que, en general, son notificadas a las personas usuarias. Estas decisiones suelen estar acompañadas de mecanismos de apelación, al menos en los términos establecidos por las propias compañías, y se enmarcan en el cumplimiento de sus normas comunitarias.

Sin embargo, en los últimos años, estas formas tradicionales de moderación han sido complementadas —y en algunos casos desplazadas— por prácticas igualmente impactantes, aunque más sutiles, menos transparentes y mucho más difíciles de identificar. Ya no se trata solo de eliminación directa, sino de intervenciones sutiles y poco claras que afectan la circulación de información de interés público y otras formas de contenido generado por las personas usuarias. La Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos (RELE) ha advertido que las compañías tecnológicas deben evitar que los algoritmos y sistemas automatizados —especialmente aquellos que operan sin una supervisión humana significativa— se conviertan en una amenaza para la libertad de expresión. El riesgo es particularmente grave cuando sus decisiones imponen restricciones excesivas y desproporcionadas que afectan con mayor frecuencia a grupos históricamente marginados.[1]

El shadow banning (o, en español, 'bloqueo en la sombra') representa estas tácticas. Mientras mantienen técnicamente disponible el contenido, estas prácticas disminuyen drásticamente su visibilidad, y afectan a medios de comunicación, activistas, emprendedores y personas usuarias que, en muchos casos, ni siquiera son conscientes de esta limitación. Lo preocupante es que esta reducción de visibilidad funciona como una forma de censura silenciosa. Sin notificaciones ni explicaciones claras, voces diversas quedan efectivamente excluidas del debate público digital, lo que compromete el pluralismo informativo y el debate democrático.

El shadow banning, si bien no bloquea completamente la capacidad de expresión de los usuarios, posee el potencial de afectar significativamente cuatro dimensiones cruciales del discurso: qué tan disponible está el contenido, su visibilidad en el ecosistema digital, su accesibilidad para diferentes audiencias y, en última instancia, su capacidad de generar impacto en la conversación pública.

Esta advertencia cobra especial relevancia a la luz del criterio establecido por la Corte Interamericana de Derechos Humanos (CIDH), que ha consagrado un principio fundamental: la expresión y su difusión constituyen un todo indivisible. Esta interpretación amplía significativamente el alcance protector del derecho a la libertad de expresión, garantizando no solo la posibilidad de la manifestación de ideas y opiniones,

1 Comisión Interamericana de Derechos Humanos (CIDH), Relatoría Especial para la Libertad de Expresión. (2024). *Inclusión digital y gobernanza de contenidos en internet*. Organización de los Estados Americanos. p. 69, párr. 276. https://www.oas.org/es/cidh/expresion/informes/Inclusion_digital_esp.pdf

sino también el derecho esencial de utilizar cualquier canal apropiado para que estas ideas alcancen al mayor número posible de personas y que estos potenciales receptores puedan efectivamente acceder a dicha información.[2]

Esta transformación en la moderación plantea serias interrogantes sobre transparencia y rendición de cuentas. Cuando las plataformas reducen algorítmicamente el alcance de ciertos contenidos sin notificarlo, ¿quién supervisa estas decisiones? ¿Bajo qué criterios se implementan?

Este estudio busca desentrañar el impacto de estas prácticas en medios, periodistas, activistas y personas usuarias, explorando además hasta qué punto las plataformas son transparentes sobre estos mecanismos que, aunque invisibles, reconfiguran profundamente nuestro ecosistema informativo e impactan sobre el estado psicoemocional de las personas.

El objetivo de esta investigación es examinar el fenómeno del shadow banning en plataformas digitales, mediante la identificación de casos específicos y el análisis de su impacto en la visibilidad de medios de comunicación, voces críticas y sectores subrepresentados. Además, busca evaluar el grado de transparencia de las plataformas en relación con estas prácticas y su incidencia en la participación democrática en los espacios públicos en línea.

Esta investigación, enfocada en identificar las prácticas del shadow banning en América Latina, adoptó un método predominantemente cualitativo, basado en entrevistas semiestructuradas a diversos actores del ecosistema digital, en la revisión de los términos y condiciones de las grandes tecnológicas como Meta (Instagram y Facebook) y X (antes Twitter), así como en la revisión bibliográfica. El análisis documental de las políticas y términos de servicio nos permitió contrastar las experiencias reportadas con el marco normativo declarado por las propias plataformas, y a partir de ello identificamos discrepancias significativas entre las prácticas percibidas y las políticas explícitas. Adicionalmente, la revisión bibliográfica abarcó tanto literatura académica como documentos de trabajo de organizaciones de la sociedad civil y *think tanks*, lo que nos posibilitó contextualizar el fenómeno dentro del debate global sobre moderación de contenidos y libertad de expresión en entornos digitales.

Este método nos brindó la oportunidad de documentar y validar las experiencias de primera mano de activistas, periodistas, defensores de derechos humanos y emprendedores digitales que identificaron haber sido sometidos a restricciones de visibilidad poco transparentes en diversas plataformas. A través de estos testimonios, pudimos construir patrones comunes en las experiencias reportadas e identificar impactos concretos en el ejercicio de la libertad de expresión en entornos digitales.

Es importante señalar que la investigación del shadow banning enfrenta desafíos metodológicos considerables desde la perspectiva técnica y de recopilación de datos. Actualmente, las principales plataformas digitales han implementado restricciones sustanciales al

2 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet,* párr. 276.

acceso investigativo, ya sea mediante esquemas enredados para la solicitud de acceso a datos, la limitación en el uso de las API, el cierre de herramientas como Crowdtangle —que anteriormente facilitaban la investigación independiente— o una reducción generalizada en la transparencia de sus operaciones algorítmicas. Estas barreras técnicas dificultan enormemente la posibilidad de documentar el fenómeno del shadow banning mediante evidencia cuantitativa robusta, lo que nos llevó a priorizar la documentación cualitativa de casos representativos.

La naturaleza inherentemente opaca del shadow banning constituye quizás el mayor obstáculo metodológico. A diferencia de otras formas de moderación de contenidos en las que existe una notificación explícita, el shadow banning se caracteriza precisamente por la ausencia de transparencia y comunicación con las personas usuarias afectadas. Esta característica, sumada a la ambigüedad en la redacción de las políticas de las plataformas —que rara vez mencionan expresamente estas prácticas o utilizan eufemismos como «reducción de distribución» o «ajustes de visibilidad»—, crea un escenario donde la documentación sistemática se vuelve extraordinariamente compleja. Por ello, nuestra metodología se centró en triangular las experiencias reportadas con los cambios observables en el alcance y la visibilidad del contenido, reconociendo las limitaciones inherentes a la investigación de prácticas deliberadamente diseñadas para ser imperceptibles. Cabe destacar que la literatura académica y técnica sobre el shadow banning está principalmente disponible en inglés, siendo todavía un campo relativamente poco explorado en español.

02.

EL SHADOW BANNING: LA INVISIBILIZACIÓN OCULTA DE CONTENIDOS Y USUARIOS

En el espacio digital dominado por las grandes plataformas tecnológicas —donde millones de personas debaten, comparten y acceden a la información— opera una forma sigilosa de silenciamiento que afecta a activistas, periodistas y otras personas usuarias, aunque pocas logran detectarla a tiempo. Se trata del shadow banning, un conjunto de prácticas de moderación mediante las cuales las plataformas reducen discretamente el alcance y la visibilidad de ciertos perfiles o publicaciones, sin notificar a quien ha sido afectado. Por eso se habla de una forma de moderación «oculta» en un doble sentido: invisibiliza el contenido para el resto de las personas usuarias y oculta la sanción para quien la padece. Como un fantasma que recorre los algoritmos, el shadow banning deja a sus víctimas atrapadas en un limbo comunicacional: siguen hablando, pero su presencia se desvanece sin dejar rastro.

A diferencia de las restricciones explícitas —como la eliminación de contenidos o la suspensión de cuentas—, el shadow banning mantiene la apariencia de normalidad. Las cuentas afectadas pueden seguir publicando con aparente normalidad, pero sus contenidos son progresivamente excluidos de las conversaciones públicas: desaparecen de los resultados de búsqueda, se vuelven invisibles en los *hashtags* más utilizados o dejan de mostrarse en los *feeds* de sus propios seguidores.[3] Es, en muchos sentidos, una versión más sofisticada del silenciamiento digital.

Las manifestaciones del shadow banning son tan diversas como sutiles: desde una caída repentina en las interacciones hasta la desaparición total de los sistemas de recomendación o de los resultados de búsqueda. Un periodista que usualmente recibe cientos de comentarios puede encontrarse, de un día para otro, hablando al vacío; una activista que emplea *hashtags* relacionados con derechos humanos descubre que sus publicaciones nunca aparecen en esas búsquedas; una educadora sexual ve cómo sus contenidos informativos son filtrados por algoritmos que los consideran «contenido límite».

Lo más preocupante es que estas restricciones algorítmicas no se aplican de manera equitativa. La evidencia indica que afectan desproporcionadamente a comunidades marginadas: activistas sociales y políticos, periodistas independientes, personas LGBTQIA+ y educadores en sexualidad integral. El problema de fondo es la opacidad. Sin notificaciones, explicaciones ni mecanismos efectivos de apelación, las personas afectadas se ven forzadas a realizar un trabajo invisible y agotador: desde formular hipótesis sobre el funcionamiento del algoritmo hasta modificar su lenguaje (*algospeak*) para evitar ser penalizadas, o construir redes colectivas para verificar y sortear la censura encubierta.

Este fenómeno no es una rareza técnica, es una amenaza concreta a los derechos fundamentales de la libertad de expresión, especialmente en contextos como América Latina, donde la visibilidad digital puede ser clave para el activismo, la denuncia o el ejercicio democrático. Urge, por tanto, avanzar hacia mayores mecanismos de rendición de cuentas y transparencia sobre las decisiones que toman quienes hoy

3 Nicholas, G. (2022). *Shedding light on shadow banning*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2022/04/remediated-final-shadowbanning-final-050322-upd-ref.pdf

controlan las puertas de acceso a la información pública.

Las manifestaciones más comunes del shadow banning incluyen:

- **Reducción del alcance de las publicaciones del usuario:** Los contenidos que una persona publica llegan a una audiencia significativamente menor de lo habitual, sin haber sido eliminados. Por ejemplo, una publicación que normalmente recibiría cientos o miles de interacciones («me gusta», comentarios o compartidos) obtiene muchas menos porque la plataforma decide no mostrarla en los feeds de otros usuarios o no la jerarquiza lo suficiente para que lleguen a verla. La ubica más abajo —o al final—, lo que en la práctica implica que alguien la vea. Esta reducción puede ser repentina o gradual en los «me gusta», comentarios y visualizaciones, de forma desproporcionada con respecto al tamaño de seguidores y métricas típicas de interacción. Algunos autores han documentado que «las publicaciones de aquellos que reportaron ser [bloqueados en la sombra] no aparecen en los feeds de sus seguidores en absoluto, y aparentemente son despriorizadas del algoritmo por completo».[4]

- **Restricción de la visibilidad del usuario en los resultados de búsqueda:** Aunque un usuario escriba exactamente el nombre de la cuenta en el buscador, esta no aparece entre los resultados, no es sugerida[5] o aparece muy abajo, lo que dificulta su localización. Esto limita el crecimiento orgánico de la cuenta y su participación en debates públicos.

- **Eliminación de la cuenta de las sugerencias y recomendaciones visibles para otros usuarios:** La plataforma deja de incluir a esa cuenta en secciones como «Personas que quizás conozcas», «Sugerencias para ti», «Cuentas recomendadas» o similares. Por ejemplo, una cuenta deja de ser sugerida a usuarios que podrían estar interesados, lo que reduce su posibilidad de llegar a nuevas audiencias.

- **Exclusión de *hashtags*, *feeds* de descubrimiento y tendencias:** Aunque la persona etiquete su publicación con un *hashtag* específico, esta no aparece cuando otros usuarios hacen clic en ese *hashtag*. Por ejemplo, un usuario que publica bajo #FreePalestine o #MeToo puede notar que su post no figura entre los resultados de esa etiqueta, lo que impide que su mensaje se sume a la conversación pública. También se refiere a que las publicaciones no aparezcan en páginas de descubrimiento algorítmico, como la página «Explorar» de Instagram o la página «Para ti» de TikTok.

- **Limitación en la interacción con otros usuarios:** Los comentarios o respuestas del usuario aparecen ocultos o relegados para otros, aunque pueden ser visibles para el propio autor. Por ejemplo, un periodista comenta una publicación viral, pero su comentario es invisible para el resto,

4 Blunt, D., Wolf, A., Coombes, E., y Mullin, S. (2020). P*osting into the void: studying the impact of shadowbanning on sex workers and activists.* Hacking//Hustling.

5 Le Merrer, E., Morgan, B. y Trédan, G. (2021). Setting the record straighter on shadow banning. En I*EEE Infocom 2021-IEEE Conference on Computer Communications* (pp. 1-10). IEEE. https://arxiv.org/pdf/2012.05101

reduciendo así su capacidad de participar en conversaciones públicas.

- **Bloqueo de características:** Imposibilidad de utilizar ciertas funciones que permiten la interacción con otras personas usuarias. Por ejemplo, no tener la capacidad de dar «me gusta», responder a publicaciones de otros usuarios o que una publicación no esté vinculada al nombre de la persona.[6]

Todas estas acciones no solo hacen que sus contenidos tengan una circulación o alcance reducido, sino que además reducen o impiden la «descubribilidad» tanto de sus publicaciones como de su propia cuenta y presencia en la red, lo que dificulta el crecimiento de su audiencia o comunidad de seguidores.

El problema de esta práctica, implementada por las plataformas para sancionar el supuesto incumplimiento de sus normas comunitarias, es que limita la circulación de las expresiones del usuario sin que este lo perciba. A diferencia de un bloqueo abierto —en el que la cuenta es suspendida o eliminada de forma directa y, por lo general, con una notificación de la plataforma—, estas restricciones operan de manera silenciosa. Aunque su cuenta no esté bloqueada en términos estrictos, al restringir la circulación de sus expresiones y limitar su capacidad de ser descubierto por nuevas audiencias, el impacto es el mismo: se obstaculiza —o directamente se excluye— su participación en el debate público en línea.

El shadow banning se aplica de manera automatizada mediante algoritmos vinculados a la inteligencia artificial que «moderan» la circulación de contenidos para controlar el discurso dentro de la plataforma. Su falta de transparencia y la ausencia de mecanismos claros para detectarlo o apelar sus decisiones lo convierten en una de las formas más polémicas de intervención de estas empresas en los nuevos espacios públicos en Internet.

Estas prácticas afectan gravemente la libertad de expresión y el pluralismo informativo, ya que medios de comunicación, periodistas y activistas pueden ver reducido su impacto en la conversación pública sin tener mecanismos adecuados y oportunos para reclamar por sus derechos y revertir la medida.

Para comprender adecuadamente el fenómeno del shadow banning, es fundamental distinguir dos conceptos clave que operan en las plataformas digitales como X, Instagram, Facebook, entre otros: la moderación de contenidos y la curación de contenidos.

El primero comprende el conjunto de políticas, sistemas y herramientas que las plataformas implementan para gestionar el material generado por sus usuarios, determinando qué se publica, qué se elimina o cómo se controla.[7] Este proceso puede estructurarse mediante tres sistemas: 1) centralizado (como en X, Facebook o YouTube), en el que la plataforma aplica reglas internamente; 2) «distribuido» (como en Reddit

6 Blunt, D., Wolf, A., Coombes, E. y Mullin, S. (2020). *Posting into the void: studying the impact of shadowbanning on sex workers and activists.*

7 Center for Democracy & Technology. (2021). *Outside looking in: approaches to content moderation in end-to-end encrypted systems.* https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/

o Wikipedia), en el que las propias comunidades gestionan la moderación con mínima intervención de la plataforma; o 3) «híbrido» (como en Twitch), en el que se integran ambos enfoques. La moderación, en términos generales, se desarrolla a través de fases secuenciales que abarcan desde la definición de reglas hasta los mecanismos de apelación, y puede aplicarse tanto antes de la publicación (*ex ante*) como después (*ex post*).[8]

Para definir el concepto de 'moderación de contenidos en redes sociales', la Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos (RELE) adopta definiciones provenientes del proceso de Diálogo de las Américas, así como de documentos de organizaciones de la sociedad civil especializadas en la materia. Así, en el párrafo 187 del informe *Inclusión digital y gobernanza de contenidos en Internet,* publicado en junio de 2024,[9] «la moderación de contenidos se define como la práctica organizada de filtrar el contenido generado y visto por las personas usuarias y publicado en plataformas digitales». En el informe se enumeran los diferentes tipos de moderación de contenido: premoderación, posmoderación, moderación reactiva, moderación distribuida y moderación automatizada.

El Relator destaca también en su informe que «el proceso de moderación puede realizarse por medio de la acción directa de una persona o por procesos automatizados basados en herramientas de inteligencia artificial junto con el procesamiento de largas cantidades de datos de las personas usuarias».[10] La moderación puede implicar «dar de baja el contenido permanente o temporalmente, en toda la plataforma o en relación con ciertos grupos de personas usuarias en un área geográfica específica, o afectar cuentas de personas usuarias bajo distintas modalidades».[11] Otro tipo de moderación puede involucrar acciones como etiquetar contenidos, brindar información extra y contextualizada sobre una publicación o desmonetizar las publicaciones, entre otras.

La curación de contenidos, en cambio, es el proceso mediante el cual las plataformas digitales seleccionan, organizan y presentan contenidos a una audiencia, de acuerdo a criterios desconocidos para sus usuarios. Determina qué contenidos tendrán mayor visibilidad y cuáles quedan relegados en los *feeds*, los resultados de búsqueda y las recomendaciones personalizadas de los usuarios de la plataforma.

La RELE considera a la curación de contenidos «como las decisiones automatizadas sobre el alcance, clasificación, promoción o visibilidad de los contenidos. Las plataformas suelen curar los contenidos con base en las recomendaciones personalizadas para los perfiles de las personas usuarias».[12] A la vez, advierte: «En la medida en la que se privilegian ciertos contenidos, la curación puede terminar por amplificar o

8 Klonick, K. (2018). *The new governors: the people, rules, and processes governing online speech.* St. John's University School of Law. https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1184&context=faculty_publications
9 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet.*
10 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet,* párr. 187.
11 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet,* párr. 187.
12 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet,* párr. 188.

reducir el alcance de determinados discursos».[13]

En este sentido, la curación de contenidos no es neutral, ya que responde a criterios definidos por cada plataforma, influenciando lo que los usuarios pueden ver y lo que queda oculto.

Estos procesos están mayormente automatizados y gestionados por sistemas algorítmicos y de inteligencia artificial que analizan la actividad de los usuarios para decidir qué contenidos promover y cuáles limitar en función de criterios como «lograr que sea un lugar seguro para la inspiración y la expresión». Sin embargo, estos criterios responden al modelo de negocio y a los intereses comerciales de las *big techs,* orientados a captar la atención de los usuarios y lograr su permanencia en las plataformas. Más recientemente, los cambios en las políticas de varias de las principales plataformas digitales han confirmado que también existen decisiones políticas para definir esos criterios. El shadow banning ocupa una posición singular en el espectro de prácticas de gobernanza de contenidos, situándose en la intersección entre la moderación y la curación de contenidos. No constituye una eliminación completa del contenido (moderación tradicional), sino una intervención algorítmica que reduce significativamente su visibilidad o alcance (curación negativa).

Específicamente, el shadow banning se ubica predominantemente dentro del ámbito de la curación de contenidos, pues afecta directamente cómo se distribuye y presenta el material a otros usuarios sin eliminarlo. Sin embargo, cuando esta reducción de visibilidad se aplica como consecuencia de infracciones percibidas a las normas comunitarias, también funciona como una forma de moderación *ex post* menos severa que la eliminación completa. La característica definitoria del shadow banning —y lo que lo hace particularmente problemático desde una perspectiva de derechos humanos— es su naturaleza deliberadamente opaca: a diferencia de otras medidas de moderación, en las que la plataforma notifica al usuario sobre la acción tomada, el shadow banning opera intencionalmente sin transparencia, dejando al usuario en la incertidumbre sobre por qué su contenido no alcanza a su audiencia habitual.

13 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet*, párr. 188.

03.

ESTUDIOS Y EVOLUCIÓN CONCEPTUAL DEL SHADOW BANNING

El shadow banning ha experimentado una significativa evolución conceptual desde sus orígenes en los primeros foros de Internet. Inicialmente, el término se refería específicamente a una técnica de moderación mediante la cual los comentarios y las publicaciones de las personas usuarias consideradas «problemáticas» —es decir, identificadas por acosar o *trolear* a otras personas— eran ocultados para el resto de la comunidad, mientras que, para la persona afectada, se mantenía la ilusión de que su contenido seguía siendo visible.[14] Esta estrategia buscaba principalmente evitar que las personas sancionadas crearan nuevas cuentas al percibir la restricción.

Savolainen proporciona un enfoque analítico más sociocultural al buscar entender el shadow banning como un elemento del «folklore algorítmico», es decir, a partir de un conjunto de «creencias y narrativas sobre los sistemas de moderación que se transmiten informalmente y pueden existir en tensión con los relatos oficiales».[15] Esta perspectiva subraya cómo el término funciona como un punto de articulación discursiva para experiencias diversas pero conectadas de gobernanza de plataformas, unificadas por una sensación compartida de opacidad e incertidumbre. Otras investigaciones han destacado cómo este tipo de restricciones afectan desproporcionadamente a usuarios de comunidades marginadas. En particular, algunos autores sugieren que las personas usuarias han desarrollado lo que denominan «teorías algorítmicas populares»[16] para intentar descifrar cómo funcionan los algoritmos de las plataformas. Estas estrategias incluyen cambios en el uso de *hashtags*, alteración de imágenes o incluso la creación de cuentas secundarias para verificar si han sido *shadow banned*.

La investigación de Kojah y otros autores, por su parte, profundiza en esta línea al definir el shadow banning como «una controvertida forma de gobernanza de plataformas caracterizada por el uso de algoritmos opacos para reducir o degradar contenido».[17] Su análisis caracteriza esta práctica como una forma de censura «insidiosa pero ligera» que impacta múltiples dimensiones de la experiencia de las personas usuarias, desde la visibilidad y las ganancias hasta la salud mental y las comunicaciones interpersonales.

14 Cole, S. (31 de julio de 2018). *Where Did the Concept of «Shadow Banning» Come From?* VICE. https://www.vice.com/en/article/where-did-shadow-banning-come-from-trump-republicans-shadowbanned/

15 Savolainen, L. (2022). Algorithmic lore and the myths of non-promotion. *Information*, *Communication & Society, 25*(8), p. 1096.

16 Karizat, N., Delmonaco, D., Eslami, M.y Andalibi, N. (18 de octubre de 2021). Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction 5*,(CSCW2): 1-44. https://dl.acm.org/doi/10.1145/3476046

17 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. y Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. *Proceedings of the ACM on Human-Computer Interaction, 9*(1), 1-22. p. 19. https://hdl.handle.net/2027.42/196532

Características transversales del shadow banning

Independientemente de la manifestación específica, los estudios coinciden en identificar ciertas características definitorias del shadow banning:

1. **Opacidad**: Ausencia de notificación o explicación de la plataforma sobre la restricción.
2. **Gradación variable**: El shadow banning no es binario, sino que existe en un continuo de reducción de visibilidad.
3. **Acumulación de efectos**: Diferentes tipos de *shadow ban* pueden coexistir, amplificando su impacto.
4. **Detección indirecta**: Las personas usuarias desarrollan métodos informales para verificar si están siendo «*shadow banneadas*», como comparar su visibilidad con la de otras personas usuarias o utilizar herramientas de terceros.

La tipología descrita previamente, sumada a las características transversales, refleja la complejidad y evolución de las prácticas de moderación algorítmica en plataformas digitales. Y muestra cómo la diversificación del término shadow banning, más allá de su definición original, refleja cómo las personas usuarias conceptualizan y responden a formas emergentes de gobernanza algorítmica caracterizadas por la opacidad y la incertidumbre.[18]

Impactos desiguales: grupos afectados y dinámicas de marginación

Un hallazgo consistente en la literatura revisada es que el shadow banning afecta desproporcionadamente a grupos ya marginados. Algunas investigaciones apuntan a que este tipo de moderación es más común en contenido relacionado con la sexualidad, la identidad racial o la protesta social. Por ejemplo, Instagram ha sido criticado por su censura sobre imágenes de cuerpos femeninos, incluyendo las publicaciones de activistas del movimiento Free the Nipple.[19]

Algunas investigaciones han documentado específicamente cómo las políticas de contenido «en el límite» (o *borderline* en inglés) impactan negativamente a comunidades vulnerables como trabajadoras sexuales, educadoras sexuales y miembros de la comunidad LGBTQIA+. Reportes periodísticos han sugerido patrones similares respecto a personas negras,[20] mujeres[21] y comunidades Queer.[22]

18 Savolainen, L. (2022). Algorithmic Lore and the Myths of Non-Promotion. Information.
19 Are, C. (2021). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies, 22*(8), 2002-2019. p. 2002.
20 BBC. (2020). *Facebook and instagram to examine racist algorithms*. British Broadcasting Corporation. https://www.bbc.com/news/technology-53498685
21 Cook, J. (2019). *Instagram's shadow ban on vaguely 'inappropriate' content is plainly sexist.* Huffington Post. https://www.huffpost.com/entry/instagram-shadowbansexist_%20n_5cc72935e4b0537911491a4f
22 Joseph, C. (2019). Instagram's murky 'shadow bans' just serve to censor marginalised communities. *The Guardian*. https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive

Un estudio de diarios y entrevistas con ocho creadores de contenido en situación de marginación documentó cómo «los creadores con identidades marginadas (mujeres, bailarines de pole dance, personas de talla grande o LGBTQIA+) se ven desproporcionadamente afectados por el shadow banning».[23] Los participantes percibían que «las personas con identidades marginadas eran más afectadas por el shadow banning y otras formas de censura que los hombres que creaban contenido similar, y que las personas que se ajustan a estándares de belleza convencionales».[24]

La investigación de Are ofrece un análisis particularmente detallado sobre cómo el shadow banning afecta a bailarines de pole dance en Instagram, revelando cómo la censura de contenido relacionado con pole dance refleja percepciones sesgadas sobre el cuerpo femenino y las expresiones de sexualidad. Su investigación destaca cómo incluso actividades artísticas y deportivas pueden ser incorrectamente categorizadas como «contenido sexualmente sugestivo»[25] cuando son realizadas por ciertos cuerpos o comunidades.

23 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. y Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 2.
24 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. y Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 9.
25 Are, C. (2021). The Shadowban Cycle. p. 2004.

El trabajo invisible de la persona usuaria bajo el shadow banning

Una contribución particularmente valiosa de la literatura reciente es la identificación del trabajo invisible que las personas usuarias deben realizar para navegar, mitigar y adaptarse a los sistemas de moderación opacos. La investigación hace un aporte significativo al identificar tres categorías específicas de trabajo invisible:

1. **Trabajo mental y emocional:** La carga cognitiva y psicológica de anticipar constantemente qué contenido podría ser restringido. Como explica un participante del estudio: «Desviar mi atención de mi creación para solucionar problemas de alineación con las políticas viene como una distracción, lo que obstaculiza mi consistencia en la producción de contenido».[26]
2. **Trabajo sin rumbo:** Esfuerzos realizados con la esperanza de evitar el shadow banning que no contribuyen directamente a la producción de contenido creativo. Esto incluye prácticas como publicar selfies «para el algoritmo» después de contenido potencialmente controvertido o con fuerte potencial de ser restringido como publicaciones pro Palestina, o utilizar algospeak[27] (modificar palabras potencialmente problemáticas).
3. **Trabajo comunitario:** La labor colaborativa entre creadores para compartir estrategias y apoyarse mutuamente. Como describe un participante en el estudio: «He trabajado con otros creadores para ayudar a impulsar el engagement durante un shadow ban. Promovemos y participamos en el contenido de cada uno durante una suspensión sospechada».[28]

26 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. y Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p.13.
27 Este término, señalan Kojah et al. (2025), se refiere a «escribir mal intencionadamente palabras o bloquear imágenes para engañar al algoritmo y eludir la moderación y supresión de contenidos, una práctica que requiere mucho tiempo y esfuerzo» (p. 14).
28 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. y Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 16.

04.

IMPACTO EN LOS MEDIOS DE COMUNICACIÓN Y EN LAS PERSONAS USUARIAS

El shadow banning opera como una forma de censura digital particularmente perniciosa: invisible, poco documentada y abrumadora para quienes señalan experimentarla. En América Latina, activistas, educadores sexuales, periodistas y pequeños emprendedores experimentan cómo sus publicaciones —que en algún momento alcanzaban miles de vistas y revisten en ocasiones asuntos de interés público— repentinamente llegan solo a unos cientos de personas, sin explicación ni advertencia. Esta reducción algorítmica del alcance no solo afecta la visibilidad de su contenido, sino que genera profundos impactos psicoemocionales caracterizados por ansiedad constante, sentimientos de impotencia y una autocensura preventiva que silencia voces en el debate público.

El creador afectado queda atrapado en un limbo digital en el que sigue publicando para una audiencia fantasma, invirtiendo tiempo y recursos en un esfuerzo que las plataformas han condenado silenciosamente a la irrelevancia.

La cuestión adquiere mayor gravedad cuando vemos que la moderación que hacen las grandes tecnológicas impacta en temas de salud sexual, diversidad corporal, cobertura periodística o crítica política, precisamente aquellos discursos que gozan de especial protección[29] en el sistema interamericano de derechos humanos.

Además, esta continua incertidumbre obliga a las personas usuarias a desplegar diferentes estrategias —a veces en solitario, otras en colectivo—, así como a invertir recursos no para mejorar sus contenidos o productos, sino para seguir en una lucha desigual contra sistemas algorítmicos opacos que condenan su presencia digital a un rincón ensombrecido.

Lo que podría considerarse una simple decisión empresarial sobre distribución de contenidos se convierte en un poderoso mecanismo de control social. Sumado a la imposibilidad de predecir qué desencadenará restricciones —desde mencionar términos como *marihuana, educación sexual, Free Palestine* hasta mostrar cuerpos no normativos—, genera un efecto inhibitorio por el cual las personas consideran preventivamente su expresión en temas de interés público.

Este «ostracismo digital» resulta especialmente grave por su opacidad.

29 La jurisprudencia del sistema interamericano de derechos humanos ha establecido tres categorías de discursos especialmente protegidos, reconociendo su papel fundamental en el fortalecimiento democrático y el ejercicio pleno de los derechos humanos. En primer lugar, la protección al discurso político y sobre asuntos de interés público, considerando que este tipo de expresiones son esenciales para la formación de una opinión pública informada y para la participación de las personas en los procesos democráticos y los asuntos públicos. En segundo lugar, la protección reforzada al discurso sobre personas servidoras públicas en el ejercicio de sus funciones y sobre personas candidatas a cargos públicos, entendiendo que el escrutinio de las acciones de quienes ostentan o aspiran a posiciones de poder es crucial para la transparencia y la rendición de cuentas. Finalmente, la protección especial al discurso que constituye un elemento de la identidad o dignidad personal del emisor, reconociendo así la importancia de la libertad de expresión para el desarrollo individual y la autonomía personal. Dicha protección diferenciada busca garantizar que esos tipos de discursos, fundamentales para el debate público y la realización personal, no sean indebidamente restringidos, promoviendo así una sociedad más abierta, plural y democrática. Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos [CIDH]. (2009). *Marco jurídico interamericano del derecho a la libertad de expresión* (párr. 32-56). Organización de los Estados Americanos. https://www.oas.org/es/cidh/expresion/docs/publicaciones/MARCO%20JURIDICO%20INTERAMERICANO%20DEL%20DERECHO%20A%20LA%20LIBERTAD%20DE%20EXPRESION%20ESP%20FINAL%20portada.doc.pdf

Las personas afectadas continúan produciendo contenido que prácticamente nadie ve, experimentando un aislamiento que deteriora tanto sus proyectos profesionales como su bienestar emocional.

Desde la perspectiva del derecho internacional de los derechos humanos, el shadow banning constituye una forma particularmente problemática de restricción a la libertad de expresión. Al analizar esta práctica bajo el prisma del artículo 19 del Pacto Internacional de Derechos Civiles y Políticos (PIDCP) o el artículo 13 de la Convención Americana, surge un incumplimiento de los requisitos fundamentales de legalidad, necesidad y proporcionalidad que deben satisfacer todas las limitaciones legítimas a este derecho. La legalidad se ve comprometida porque estas restricciones se implementan mediante términos de servicio ambiguos y algoritmos opacos que no proporcionan certeza jurídica sobre qué expresiones pueden ser limitadas. Tampoco se cumple con los principios de necesidad y proporcionalidad cuando las plataformas aplican estas medidas de forma automatizada, sin evaluar el contexto específico ni el impacto en el debate público, y sin considerar alternativas menos restrictivas, como las advertencias o etiquetas.

La Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos (RELE) ha establecido claramente que cualquier restricción debe ser específica y aplicada mediante decisiones motivadas que permitan una responsabilidad posterior. El shadow banning, por definición, viola estos principios al implementar restricciones sin notificación, explicación ni posibilidad efectiva de impugnación. También se hace sobre la base de unos términos de servicio confusos e imprecisos, que abre la interrogante sobre si el shadow banning, además de incumplir con el principio de legalidad, es una forma de censura previa y una restricción al derecho a la libertad de expresión a través de medios indirectos. La RELE ha señalado que este derecho «no puede ser objeto de medidas de control preventivo o previo, sino de la imposición de responsabilidades posteriores para quien haya abusado de su ejercicio».[30]

La responsabilidad de las grandes plataformas tecnológicas frente al shadow banning es ineludible. Los principios rectores sobre las empresas y los derechos humanos de la ONU (UNGP por sus siglas en inglés) establecen que las empresas tienen la responsabilidad de respetar los derechos humanos, independientemente de la capacidad o voluntad de los Estados de cumplir sus obligaciones.[31] Para los gigantes tecnológicos como Meta, TikTok o X, esto implica tres obligaciones concretas: 1) identificar y prevenir impactos negativos en derechos humanos, 2) implementar procesos apropiados según su escala y 3) proporcionar mecanismos efectivos de reparación para las víctimas.

30 CIDH. (2009). *Marco jurídico interamericano del derecho a la libertad de expresión,* párr. 91, p. 31.
31 Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. (2011). *Principios rectores sobre las empresas y los derechos humanos: Puesta en práctica del marco de las Naciones Unidas para «proteger, respetar y remediar».* Principio 11, p. 15. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_sp.pdf

La RELE ha sido enfática en señalar que

> [el] ejercicio del poder normativo de moderación por parte de las plataformas de internet, sobre todo las grandes plataformas, debe alinearse con los principios de los derechos humanos, el fomento del debate público y la consolidación de la democracia en las Américas. No solo deben adherirse a las normas del sistema interamericano, sino también ajustar su poder a estándares de transparencia y rendición de cuentas, basados en la igualdad y la no discriminación. Esto es crucial para crear un ambiente en línea que respete los derechos humanos y sea libre, abierto e inclusivo, y que fomente la autonomía y los derechos de los usuarios.[32]

La brecha entre el discurso corporativo y la realidad resulta preocupante. Mientras Adam Mosseri de Instagram afirmaba en 2021 que «si algo hace tu contenido menos visible, debes saberlo y poder apelar», la experiencia cotidiana de creadores, activistas y periodistas latinoamericanos revela la ausencia sistemática de notificaciones y mecanismos efectivos de revisión cuando su contenido es restringido.

Esta contradicción subraya la urgencia de marcos regulatorios que reconozcan a las plataformas digitales como espacios de debate público donde se desarrollan las discusiones democráticas contemporáneas. En una región donde la concentración mediática históricamente ha limitado la pluralidad de voces, las redes sociales inicialmente representaron una promesa democratizadora. Sin embargo, esta se ve hoy comprometida por prácticas de moderación opacas que reproducen —e incluso amplifican— exclusiones preexistentes, aunque lo hagan bajo el velo de decisiones algorítmicas aparentemente neutrales y tecnocráticas.

Lo que está en juego, además de las métricas y el alcance, es el derecho de las sociedades democráticas a un ecosistema informativo diverso, plural y accesible en el que las voces tradicionalmente marginadas puedan participar en igualdad de condiciones en la construcción del debate público latinoamericano.

La paradoja resulta aún más evidente al contrastar las promesas corporativas con los términos y condiciones de las plataformas. Mientras X menciona explícitamente en sus políticas que «limitará la visibilidad» de ciertos contenidos y afirma que los usuarios afectados recibirán notificaciones y podrán solicitar revisión, lo recabado en esta investigación apunta a que estas garantías rara vez se materializan. Meta, por su parte, admite abiertamente que reduce la distribución de «contenido problemático» sin comprometerse a informar a las personas afectadas o proporcionarles un mecanismo de apelación claro.

¿Qué ocurre cuando las garantías prometidas en términos y condiciones se convierten en letra muerta frente a la experiencia real de los usuarios? ¿Cómo pueden las personas en un contexto como Latinoamérica defender sus derechos cuando ni siquiera saben que están siendo restringidos? ¿Quién supervisa

32 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet,* p. 61, párr. 264.

que las decisiones algorítmicas no reproduzcan sesgos contra voces históricamente marginadas? Estas preguntas permanecen abiertas mientras las personas navegan a ciegas en un ecosistema digital en el que las reglas escritas rara vez coinciden con las prácticas aplicadas, y donde la promesa de una mayor libertad de expresión cede terreno frente a una nueva forma de exclusión digital tan efectiva como invisible.

El debate público robusto y pluralista, esencial en una sociedad democrática, requiere que tanto los Estados como las grandes plataformas tecnológicas aborden esta forma invisible de silenciamiento. Las personas deben poder conocer cuándo y por qué su contenido es restringido, contar con mecanismos efectivos para impugnar estas decisiones y tener garantías de que las limitaciones a su libertad de expresión cumplirán con los estándares internacionales de derechos humanos. Sin estas salvaguardas, el shadow banning continuará socavando silenciosamente la vitalidad del debate democrático en la región, creando la ilusión de participación mientras se ejercen formas sofisticadas de control sobre el discurso público.

05.

ALGUNOS EJEMPLOS DE CASOS DE SHADOW BANNING

Con el objetivo de comprender en profundidad cómo opera el shadow banning en las plataformas digitales, esta investigación incluye el análisis de una serie de casos representativos de América Latina y el sur global. Se trata de situaciones en las que periodistas, activistas y proyectos de comunicación e incidencia social han denunciado una reducción significativa del alcance, la visibilidad o la «descubribilidad» de sus contenidos y de sus cuentas, sin recibir explicaciones claras ni notificaciones por parte de las plataformas.

Uno de los principales ejes de análisis será el tratamiento que Facebook, Instagram, y X han dado a las cuentas y publicaciones que difunden información en apoyo a la causa palestina, en el marco de la ofensiva militar de Israel contra la población de ese territorio, ampliamente denunciada como un genocidio por parte de expertos, gobiernos, agencias de la ONU y la Corte Penal Internacional.[33] Diversas denuncias públicas y estudios han documentado la aplicación de mecanismos de moderación opacos que limitan el alcance de estos contenidos, lo que representa una amenaza al pluralismo informativo y al derecho a expresarse libremente en entornos digitales. En el marco de esta investigación, entrevistamos a Mona Shtaya, Campaigns and Partnerships Manager (MENA) y Corporate Engagement Lead en Digital Action, quien utiliza su cuenta de Instagram para difundir información sobre la situación en Palestina y ha denunciado públicamente las restricciones que enfrentan estos contenidos en las plataformas digitales.

Otro caso documentado es el del medio mexicano Chiapas Sin Censura, que sufrió una drástica reducción de alcance en Facebook tras recibir una infracción por «fraude» asociada a una publicación de años anteriores. La sanción —sin previo aviso ni posibilidad efectiva de apelación— afectó tanto la visibilidad como la monetización del medio, y forzó a su equipo a limitar la publicación de ciertos contenidos por temor a nuevas penalizaciones.

Junto a este caso, se relevarán también otras experiencias significativas, como la de la activista cannábica chilena conocida por los proyectos Santiago Verde y Muy Paola, quien reportó reiteradas restricciones de visibilidad en Instagram. En Perú, la cuenta Emma y Yo, un espacio de educación sexual liderado por Alesia, también denunció una caída abrupta en el alcance de sus publicaciones, especialmente aquellas vinculadas a temas de derechos sexuales y reproductivos.

Asimismo, se analizarán casos ocurridos en Argentina, como el de la fotoperiodista y doctora en Ciencias Sociales Cora Gamarnik, quien evidenció reducciones drásticas de alcance en Facebook sin que la plataforma ofreciera mecanismos para apelar la decisión; y el del periodista Sebastián Lacunza, quien denunció en X una baja inexplicable en el alcance de sus publicaciones tras reportar sobre cuestiones políticas y de medios.

Estos casos, contextualizados en distintos marcos temáticos y geográficos, permiten visibilizar cómo el shadow banning afecta el derecho a la libertad

33 International Criminal Court. (n.d.). Palestine. https://www.icc-cpi.int/palestine; Naciones Unidas. (26 marzo de 2024). Relatora acusa a Israel de estar cometiendo un genocidio en Gaza. https://news.un.org/es/story/2024/03/1528636

de expresión y la circulación de voces diversas, particularmente cuando se trata de discursos críticos o provenientes de sectores históricamente marginados.

Shadow banning en Palestina: censura algorítmica contra un pueblo

Desde que comenzó el ataque de Israel en Gaza, diversos usuarios vienen denunciando a Instagram, Facebook, TikTok y X por limitar la visibilidad de sus publicaciones a favor de Palestina, aunque no las eliminan.[34] En las distintas redes pueden encontrarse videos de usuarios que publican contenidos a favor de Palestina, en los que cuentan que no tienen permitido transmitir en directo, que identifican una caída drástica de las interacciones o del número de reproducciones de sus videos, o que reciben mensajes de otros usuarios que les avisan que no pueden dejarles comentarios en sus publicaciones.

«Autores, activistas, periodistas, cineastas y usuarios han dicho que las plataformas están ocultando publicaciones que contienen *hashtags* como #FreePalestine y #IStandWithPalestine, así como mensajes que expresan apoyo a los civiles palestinos asesinados por las fuerzas israelíes».[35]

«La empresa dice que no hay sesgo, que todo el mundo se vio afectado, pero no es así. No hemos visto a ningún usuario israelí quejarse del *shadow ban,* ni siquiera en cuanto a las tendencias o al bloqueo de comentarios», asegura Nadim Nashif, fundador de 7amleh, también conocido como el Centro Árabe para el Avance de las Redes Sociales.[36] Para burlar el shadow banning, algunos intentan engañar a los algoritmos sustituyendo ciertos términos considerados propalestinos por otros en sus *hashtags* o publicaciones. En el caso de Meta, la censura ejercida sobre contenidos favorables a Palestina fue incluso motivo de reclamo por parte de casi doscientos empleados de la empresa a través de una carta abierta dirigida a Mark Zuckerberg en diciembre de 2023.[37]

En noviembre de 2024, el periodista palestino Younis Tirawi, conocido por exponer los crímenes de guerra israelíes, empezó a recibir mensajes de diversos usuarios de X que le informaban de «fallas» al intentar seguirlo en esa red social.[38] El medio de comunicación Decensored News mostró la «falla» mediante una grabación de pantalla, e informó que Tirawi «de repente perdió a la mayoría de sus seguidores». Muchos de

34 Observacom. (2 de octubre de 2024). Denuncian shadowban de grandes plataformas a contenidos sobre Palestina. https://www.observacom.org/denuncian-shadowban-de-grandes-plataformas-a-contenidos-sobre-palestina/
35 Observacom. Denuncian shadowban de grandes plataformas.
36 France 24 Español. (2 de noviembre de 2023). Usuarios propalestinos denuncian 'shadowban' en plataformas de redes sociales [Video]. YouTube. https://www.youtube.com/watch?v=xm9llqJjg1A&ab_channel=FRANCE24Espa%C3%B1ol
37 *Dear Mark Zuckerberg and Leadership*. https://metastopcensoringpalestine.com/
38 Observacom. (28 de noviembre de 2024). ¿Fallas o censura oculta? polémica por restricciones a periodista palestino en X. https://www.observacom.org/fallas-o-censura-oculta-polemica-por-restricciones-a-periodista-palestino-en-x/

ellos aseguraban haberlo dejado de seguir «involuntariamente y que X no les permitía volver a seguirlo»: al hacer clic en el botón «seguir», automáticamente volvían a figurar como no seguidores.

A pesar de que desde X aseguraron que fue un «inconveniente», es altamente significativo que la misma situación le había sucedido un mes antes a la cuenta @Palestinahoy01 de la misma plataforma. Durante siete días, la red social de Elon Musk no permitía a los usuarios seguir la cuenta y su número de seguidores cambiaba constantemente.

La situación se repite en palabras de Mona Shtaya —especialista palestina en derechos digitales y moderación de contenidos, actual Campaigns and Partnerships Manager para Medio Oriente y el Norte de África y Corporate Engagement Lead en la organización Digital Action—, que relató su experiencia como usuaria de Instagram afectada por los mecanismos de shadow banning: «Mi cuenta tiene más de veinte mil seguidores y se enfoca principalmente en la rendición de cuentas de las *big tech* y los derechos digitales. Desde hace dos años, he estado enfocada intensamente en Palestina debido a la situación allí».[39]

Según explicó Mona durante la entrevista, la primera vez que se documentaron casos de shadow banning en Palestina fue en mayo de 2021, durante los hechos en los que el ejército israelí intentó desplazar por la fuerza a familias palestinas en los barrios de Sheikh Jarrah y Silwan, en Jerusalén. En ese mismo período, también se produjo una ofensiva de doce días sobre la Franja de Gaza. En ese contexto, comenzaron a reportarse restricciones de visibilidad de contenidos relacionados con Palestina en las plataformas digitales.

En cuanto a su propia cuenta de Instagram, Mona señaló que el shadow banning se activó en noviembre de 2023, aproximadamente un mes después del inicio de la actual ofensiva militar de Israel sobre Gaza. Hasta ese momento, no había sufrido restricciones similares.

Mona identificó que algo estaba pasando en su cuenta de Instagram cuando empezó a comparar los números de sus estadísticas. Por ejemplo, cuando publica una *selfie*, dice, sus historias pueden llegar a más de dos mil seiscientas vistas, pero si comparte «una historia criticando a Meta por silenciar a los palestinos, esa historia no llega a doscientas vistas». Una caída de visibilidad de más del 90 %. Mona usó distintas estrategias para hacer pruebas de alcance e intentar comprender lo que estaba ocurriendo. Expresa:

> Hice colaboraciones con cuentas grandes, de más de diez millones de seguidores. Esa semana, mis publicaciones alcanzaron más de un millón de cuentas, pero mis historias apenas llegaron a doscientas vistas. Eso demuestra que hay un problema. No tiene sentido que tenga un alcance de un millón y tan poca visibilidad en las historias.

En algunos casos, las personas que intentaron buscar la cuenta de Instagram de Mona no pudieron encontrarla.

39 Entrevista realizada por Observacom y Digital Action (2025) a Mona Shtaya, 3 de abril de 2025.

También recibieron un mensaje de advertencia al intentar enviarle un mensaje privado: «¿Estás segura de que quieres enviarle un mensaje a esta persona?», preguntaba la plataforma al momento de hacer clic en «enviar».

Mona Shtaya explicó que la aplicación de shadow banning sobre su cuenta tuvo un momento claro de inicio:

> Yo no estaba siendo *shadow banneada* al comienzo del genocidio. Hice un video que se volvió viral y tuvo cerca de doscientas mil vistas. Básicamente criticaba la complicidad de Meta en el genocidio. Ese video se volvió viral y todo lo que vino después fue una locura. Todo lo que publiqué después fue fuertemente *shadow banneado*.

Antes de ese momento, Mona había trabajado con defensoras de derechos humanos que enfrentaban restricciones similares, pero nunca había experimentado esa situación en su propia cuenta.

Relata:

> Fue la primera vez que me pasó. Y después de eso, sentí que cualquier contenido que compartía tenía el mismo problema. Porque durante el genocidio, básicamente, no compartí otra cosa que contenido relacionado con el genocidio y con los derechos digitales. Y todo ese contenido claramente tenía dificultades para llegar a la gente.

En su testimonio, Mona Shtaya compartió observaciones propias y también de otras personas usuarias que documentaron casos de shadow banning en plataformas de Meta, vinculados a determinadas palabras clave y símbolos asociados a la causa palestina. Aunque en su caso no solía utilizar *hashtags* en las historias, explicó que muchas personas que sí lo hacían experimentaron una notoria reducción en el alcance de sus publicaciones cuando incluían términos como *Palestina, Gaza, genocidio* o *apartheid*.

Una de las evidencias más significativas que se conocieron sobre este tipo de prácticas fue publicada por *The Intercept* en octubre de 2024. Según esa investigación, Meta habría aplicado restricciones de visibilidad a publicaciones que incluían el triángulo rojo, símbolo que muchas personas usuarias comenzaron a utilizar para representar el apoyo a Palestina, sin que la plataforma informara de manera explícita sobre esa política.[40] Mona señaló que, aunque ya existían sospechas sobre este comportamiento, esa publicación sirvió como prueba concreta de la aplicación de medidas de reducción de visibilidad.

Otro patrón detectado tuvo que ver con los comentarios en Instagram que incluían la bandera de Palestina, corazones con los colores de la bandera o frases como «Free Palestine». Estos comentarios aparecían ocultos o marcados como «comentario oculto», sin que mediara una advertencia clara de por qué. A raíz de una investigación solicitada por la propia Mona, *The Intercept* replicó el problema en su cuenta, y al consultar

40 The Intercept. (2 de octubre de 2024). Facebook and Instagram Restrict the Use of the Red Triangle Emoji Over Hamas Association. The Intercept. https://theintercept.com/2024/10/02/meta-facebook-instagram-red-triangle-emoji/

a Meta, la empresa respondió que solo ocultaban comentarios cuando detectaban «discurso hostil». Esa explicación fue interpretada por activistas como una señal de que, al menos en la práctica, Meta estaba clasificando términos y símbolos vinculados a Palestina como contenido hostil, aunque sin admitirlo públicamente ni justificarlo con claridad.

La plataforma con el mayor porcentaje de usuarios que denuncian haber sufrido shadow banning, según una encuesta realizada por el Center for Democracy and Technology en 2022, es Facebook (8,1 %), seguida de la actual X (4,1 %), Instagram (3,8 %) y TikTok (3,2 %). Esta censura opaca tiende a afectar de forma más frecuente y dura a ciertos movimientos sociales. Además de lo que está sucediendo en el caso de usuarios que publican contenidos a favor de Palestina, esta situación se repite con la comunidad negra, el movimiento Black Lives Matter y la comunidad LGBTQIA+.

¿Por qué las plataformas deberían atender los reclamos sobre la censura que ejercen? De acuerdo con los principios rectores sobre las empresas y los derechos humanos de las Naciones Unidas (PRNU),[41] las empresas tienen la responsabilidad de evitar infringir los derechos humanos, identificar y abordar los impactos en derechos humanos de sus operaciones y proporcionar acceso a un remedio significativo para aquellos cuyos derechos han sido vulnerados.

Para las empresas de redes sociales, esto incluye alinear sus políticas de moderación de contenido con los estándares internacionales de derechos humanos, de modo que las decisiones de eliminación de contenido sean transparentes y no excesivamente amplias o sesgadas, y que sus políticas se apliquen de manera consistente. Pese a que Meta permite una cantidad significativa de expresión propalestina, esto no justifica las restricciones indebidas sobre el contenido pacífico de apoyo a Palestina, contrarias a los derechos universales de libertad de expresión y acceso a la información.

Sebastián Lacunza: penalización invisible en X

Entre fines de 2022 y septiembre de 2023, el periodista argentino Sebastián Lacunza comenzó a notar una caída abrupta en las interacciones de su cuenta en la red social X. Las estadísticas que solía consultar de forma ocasional mostraban números inusualmente bajos: disminución permanente de seguidores y escasas reacciones, incluso cuando otros usuarios con muchos seguidores lo citaban.

«Había una caída muy abrupta de todas las estadísticas y tenía una pérdida de seguidores permanente. Incluso notaba que a veces me citaba gente

41 Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. (2011). *Principios rectores sobre las empresas y los derechos humanos: Puesta en práctica del marco de las Naciones Unidas para «proteger, respetar y remediar».* https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_sp.pdf

con muchos seguidores, y como que eso igualmente seguía siendo dificultado, lo cual era raro»,[42] relató durante la entrevista.

Lacunza supo después que ese fenómeno podía estar relacionado con una práctica conocida como shadow banning, aunque en ese momento todavía no conocía el término. Recién hacia mediados de 2023, y tras escucharlo de parte de colegas o activistas, entendió que lo que estaba viviendo podía no ser una simple baja de interacciones, sino una penalización encubierta aplicada por la propia plataforma.

Desorientado, Lacunza consideró pagar la suscripción premium, aunque no tenía certeza de que eso fuera a resolver el problema. Finalmente decidió publicar un tuit denunciando la situación.[43] «Hice el tuit, y a las 48 horas lo levantaron. O menos de 48 horas». En menos de 48 horas después de haber publicado el tuit en el que denunciaba el shadow banning, su visibilidad comenzó a normalizarse. «Fue inmediato», aseguró. No hubo ninguna notificación oficial por parte de X ni una respuesta automática que explicara qué había sucedido. Simplemente, las interacciones empezaron a aumentar y las estadísticas, que habían estado en caída durante meses, comenzaron a revertirse.

«Al principio la recuperación fue bastante acelerada: ganaba entre quinientos y mil seguidores por mes. Después se estabilizó, pero algo cambió claramente», señaló. En retrospectiva, Lacunza reconoce que la sanción invisible que vivió solo pudo ser comprobada mediante observación propia y comparaciones de alcance. «Durante los meses que estuve *shadow banneado* yo estaba en un estándar que si tenía quince «me gusta» era mucho. Y de repente, después de publicar mi tuit, llegué a tener incluso una publicación con veintiún mil «me gusta». Nunca en mi vida había tenido eso», explicó. A pesar de esa mejoría, sigue sin saber si las limitaciones que sufrió en su cuenta se eliminaron por completo o si persiste algún tipo de restricción parcial.

Otro indicio claro del shadow banning fue la pérdida de «descubribilidad» de su cuenta. Sebastián Lacunza relató que, durante el período en que sospechaba estar afectado por una penalización algorítmica, resultaba notoriamente difícil encontrar su perfil incluso escribiendo su nombre y apellido completos en el buscador de la plataforma. «Yo ponía el buscador y no aparecía mi cuenta. En cambio, sí aparecían tres o cuatro cuentas falsas que alguien creó usando mi nombre y mi avatar», explicó. Esta experiencia coincidía con uno de los indicadores que otros usuarios mencionaban en foros o tutoriales para detectar un shadow banning: la exclusión de los resultados de búsqueda pese a buscar el nombre exacto del usuario.

Esta falta de visibilidad no se debía a un error del sistema ni a una baja general del motor de búsqueda, sino que afectaba exclusivamente a su cuenta, mientras perfiles falsos con su nombre sí eran fáciles de hallar. «Es un dato importante, porque no es que no aparecía ningún resultado: aparecían esas cuentas falsas, pero no la mía», señaló Lacunza.

42 Entrevista realizada por Observacom y Digital Action (2025) a Sebastián Lacunza, 18 de marzo de 2025.
43 Lacunza, S. (15 de setiembre de 2023). *Me percato que tengo shadowban de Twitter hace meses.* [Posteo]. X. https://x.com/sebalacunza/status/1702735419777880281

Su caso ilustra de forma clara cómo el shadow banning puede operar silenciosamente, sin que los usuarios tengan conocimiento de su existencia ni acceso a mecanismos claros de apelación o revisión. Y, al mismo tiempo, evidencia cómo esa sanción puede ser levantada sin explicación, lo que refuerza el carácter opaco de estas prácticas y la falta de garantías para quienes ven afectada su visibilidad en entornos digitales que hoy funcionan como espacios públicos de deliberación y expresión.

Shadow banning al medio independiente Chiapas Sin Censura

El medio mexicano Chiapas Sin Censura, fundado en 2012, ha sido un caso relevante de shadow banning, en especial en Facebook. Su fundador y director general, José David Morales Gómez, ha denunciado graves restricciones de visibilidad en redes sociales, en particular en Facebook. De acuerdo a lo que relata David, el medio sufrió una drástica caída en su alcance tras recibir una sanción por una presunta infracción vinculada a una publicación realizada cuatro años antes. La plataforma clasificó el contenido como «fraude», sin ofrecer mayores explicaciones ni la posibilidad de apelar de manera efectiva. La publicación en cuestión difundía el pedido de un joven con cáncer que solicitaba apoyo al boxeador Canelo Álvarez, conocido por sus acciones solidarias.

«Esta publicación, que era real, estaba circulando también en otras páginas y no había pasado nada. De pronto nos cae la infracción por fraude, y nos manda directamente de verde a rojo», relató David. Hasta ese momento, Chiapas Sin Censura acumulaba más de cien millones de visualizaciones mensuales; tras la penalización, el alcance se redujo a veintiocho millones. «Nuestro crecimiento orgánico se detuvo, los seguidores comenzaron a bajar, perdimos la monetización y no obtuvimos ninguna respuesta a nuestras apelaciones», explicó.

Luego de recibir la infracción, el medio intentó apelar la decisión por distintas vías, sin obtener respuestas. David explicó que utilizó la opción de apelación que la propia plataforma habilita al momento de notificar la sanción: «Decía que nos iban a responder en cuatro días hábiles, pero nunca contestaron». También envió correos electrónicos, abrió reportes manuales desde el perfil de administrador y se suscribió al servicio de Meta Verify —una herramienta paga que promete atención personalizada—, con la esperanza de obtener una respuesta más ágil. La suscripción costaba alrededor de siete mil pesos mexicanos por mes.

Durante ese proceso, logró hablar una vez con una persona de Meta, quien le indicó que el caso ya había sido remitido y que recibiría una respuesta en 24 horas. Esa fue la única interacción directa que tuvo. A partir de allí, no hubo más contacto. «Intentaba hacer otro reporte y me decía que ya había una conversación iniciada. No se pudo hacer nada más», relató.

La falta de vías efectivas para apelar fue uno de los aspectos más frustrantes del proceso. La sanción se mantuvo activa durante semanas, afectando directamente el alcance, la monetización y el vínculo con la audiencia. «Si no fuera por el apoyo de una organización que intervino para ayudarnos, yo creo que la sanción hubiera durado un año», afirmó David, en referencia al acompañamiento recibido.

Este caso pone en evidencia las limitaciones de los canales internos de apelación de Meta, lo que afecta directamente el derecho de sus usuarios a defenderse. La falta de una instancia clara de revisión —accesible, transparente y con tiempos razonables de respuesta— constituye una vulnerabilidad crítica para medios que dependen casi exclusivamente de plataformas digitales para su distribución y supervivencia. Esto también impacta negativamente en el derecho a recibir información de la ciudadanía que se informa a través de estos medios.

El impacto no fue solo técnico o económico: también afectó el trabajo periodístico. Según contó David, comenzaron a evitar publicar temas sensibles —como casos de violencia, niñas y niños en situación de vulnerabilidad o denuncias contra el crimen organizado— por temor a nuevas sanciones. «A veces decimos: esta nota no vale una infracción. Aunque el hecho sea real, preferimos no publicarlo, porque nos puede perjudicar más».

El bloqueo tuvo además un efecto emocional y editorial. «Pasé días sin ganas de publicar. ¿Para qué hacerlo si el trabajo no le llega a la gente?», expresó. Como en muchos otros medios digitales de la región, el canal principal de Chiapas Sin Censura es Facebook. «Si el día de mañana deciden desaparecer nuestra página, perdemos doce años de trabajo y diez familias se quedan sin ingresos».

Aunque finalmente la infracción fue retirada, el episodio expone la falta de mecanismos accesibles de defensa frente a decisiones automatizadas y unilaterales. También deja en evidencia cómo la amenaza de invisibilización impacta de forma directa sobre la cobertura de temas de interés público y la sustentabilidad de medios locales y críticos.

@muypaola: cuando hablar de cannabis te vuelve invisible en redes

Frente a su teléfono, mientras tenemos la videollamada para la entrevista, Paola Díaz muestra las estadísticas de su cuenta de Instagram: «Tu cuenta no se puede mostrar a las personas que no te siguen»,[44] lee la notificación sin mayor explicación. Con frustración, señala la brutal diferencia de alcance entre publicaciones aparentemente similares: un contenido de activismo recibe cinco mil trescientas visualizaciones, mientras otro de temática festiva alcanza setenta mil.

44 Entrevista realizada por Observacom y Digital Action (2025) a Paola Díaz, 12 de marzo de 2025.

La activista cannábica chilena, creadora de las cuentas @stgoverde[45] y @muypaola, señala la reducción sistemática de su alcance sin notificación explícita ni justificación transparente. No le bloquean directamente la cuenta, sino que la vuelve prácticamente invisible. «El *shadow ban* te empieza a mostrar cada vez menos y te muestra solo a tus seguidores», explica Paola mientras documenta cómo sus publicaciones, que antes alcanzaban cientos de miles de vistas, ahora apenas llegan a su círculo inmediato. Incluso su capacidad de ser encontrada en búsquedas ha sido severamente limitada: «Cuando me buscan tienen que poner el nombre completo, y apretar "enter", porque tampoco sale en sugerencias», añade, describiendo un sistema diseñado para que ciertos perfiles simplemente desaparezcan del radar público.

El caso de Paola ilustra cómo la intervención algorítmica de plataformas como Instagram puede operar con una doble invisibilidad: oculta el contenido al resto de usuarios y, simultáneamente, oculta este proceso de restricción al propio creador. A diferencia de una suspensión directa de la cuenta, el shadow banning carece de notificaciones formales o explicaciones sobre qué términos o temas están siendo penalizados. Paola ha identificado patrones específicos tras años de documentación, por ejemplo, mencionar términos como *marihuana*, *420* o incluso compartir información sobre la cultura cannábica y los derechos asociados activa inmediatamente filtros invisibles que reducen drásticamente su alcance. «Meta identifica palabras que no se pueden nombrar porque te van a dar *shadow ban*», revela, exponiendo un sistema de moderación temática altamente específico.

Esta moderación oculta ha tenido consecuencias devastadoras que van más allá de la mera visibilidad en línea. En septiembre de 2024, tras años de shadow banning intermitente, Instagram finalmente desactivó una de sus cuentas principales después de solicitarle verificación biométrica. Lo más increíble: «Me siguieron cobrando la suscripción [de Meta Verified]», cuenta indignada, describiendo cómo Meta continuó facturándole por su servicio de verificación en cuentas que ya no podía utilizar. Una vez más, la falta de transparencia se combina con prácticas comerciales cuestionables sin que existan mecanismos efectivos de apelación o reparación.

El impacto emocional y profesional, al igual que en el caso de Emma y Yo, ha sido profundo para Paola. «Yo ahora me meto a las redes sociales y me empieza la ansiedad», confiesa. El miedo constante al ostracismo digital y a quedar silenciada ha creado una autocensura preventiva que frustra su capacidad para educar sobre temas cruciales: «No puedo hablar de reducción de daños, hay montones de veces que me gustaría hablar de problemas del narcotráfico, [pero] no puedo prevenir a mi comunidad de riesgos reales». El temor a ser expelida de una plataforma tiene un efecto inhibitorio sobre su derecho a participar en el debate sobre el cannabis en Chile. La paradoja resulta evidente: mientras las plataformas justifican

45 Al momento de esta publicación, la cuenta @stgoverde permanece suspendida por Instagram, tras múltiples ciclos de restricción y recuperación. La plataforma no solo suspendió la cuenta, sino que previamente había removido sistemáticamente contenidos de interés público, incluyendo reportajes periodísticos sobre la cultura cannábica y material educativo producido por la activista.

sus políticas de moderación como protección contra daños, están silenciando precisamente las voces que podrían prevenir riesgos concretos relacionados con el consumo y la criminalización del cannabis.

La respuesta de Paola frente a esta situación ha sido multidimensional, combinando estrategias jurídicas y de visibilización colectiva. Junto con un estudio jurídico, presentó un recurso de protección en Chile que fue rechazado y ahora se encuentra en proceso de apelación. Simultáneamente, ha denunciado estas prácticas ante el Servicio Nacional del Consumidor (Sernac), argumentando que existe una violación de derechos del consumidor cuando Meta factura en Chile sin respetar la jurisdicción local. «Ellos están facturando en Chile, por lo tanto, tienen que trabajar bajo la jurisdicción chilena», sostiene, planteando un desafío regulatorio que trasciende las fronteras nacionales.

El caso de Paola es una muestra de una problemática global que afecta a diversas comunidades: desde activistas cannábicos hasta personas defensoras de derechos humanos en Palestina, pasando por creadoras de contenido de moda inclusiva con cuerpos no hegemónicos. Paola ha buscado formar redes transnacionales con personas afectadas desde Argentina hasta Tailandia, para evidenciar la naturaleza sistemática de esta censura selectiva. «Tengo un grupo con muchas personas de diferentes partes del mundo —Argentina, Tailandia, España, México, Uruguay— todos cannábicos, y a todos los han dado de baja», explica, describiendo cómo creadores con cientos de miles de seguidores han experimentado el mismo patrón de invisibilización progresiva seguido de eliminación completa de las redes.

Cora Gamarnik: suspensión y caída drástica del alcance en Facebook

«Lo que sentí es una drástica reducción del alcance y de las interacciones de mi cuenta de Facebook, sobre todo a partir de la publicación que compartí en referencia a lo de Lago Escondido, donde hice un *screenshot* del chat de los jueces», explica la doctora en Ciencias Sociales Cora Gamarnik. El 12 de diciembre de 2022, Gamarnik publicó un post en Facebook haciendo referencia al viaje de jueces, funcionarios públicos y miembros del Grupo Clarín, de Argentina, y al intercambio de mensajes entre ellos que se hicieron públicos esa misma semana.

Facebook eliminó su publicación y luego bloqueó la cuenta de Gamarnik durante un tiempo. «Cuando publiqué eso, me suspendieron la cuenta porque decían que yo era la que estaba distribuyendo mensajes de odio. Entonces, hice el reclamo y expliqué que mi post era en contra de los mensajes de odio que estaban en esa captura de pantalla que yo compartí», explica. Esa suspensión duró unos días.

Su post consistía en una serie de capturas de pantallas del chat en el que aparecían frases como «limpiemos un mapuche» y un texto escrito por ella

contra el racismo que suponía ese caso. «Claramente Facebook no leyó el texto. Después de eso, mis posts empezaron a tener alcances muy limitados, muy distinto a lo que sucedía antes», explica la investigadora del Conicet.

Luego de percibir un cambio en el alcance de sus publicaciones —debido a que comenzó a recibir muchas menos interacciones de lo habitual—, Gamarnik fue dejando de usar Facebook y comenzó a tener más actividad en otras redes sociales, a pesar de que nunca se alejó del todo. Cora explica:

> Hasta ese momento mi cuenta de Facebook tenía mucha difusión y hay cosas que yo sabía fehacientemente que si las publicaba se viralizaban o se compartían inmediatamente. Lo que empecé a ver es que esas mismas cosas no tenían ningún resultado.

Además, le pasó que muchas personas que la seguían y leían sus publicaciones con regularidad empezaron a decirle que habían dejado de verla en Facebook y que incluso habían pensado que ya no publicaba más contenidos en esa red social.

En la actualidad, si se presta atención a la evolución de sus métricas, se puede ver cómo fueron disminuyendo las interacciones de las publicaciones de Cora Gamarnik a partir del 12 de diciembre de 2022, al punto de pasar de publicaciones sobre asuntos similares con más de dos mil «me gusta», varias decenas de «comentarios» y más de trescientos «compartidos» (a comienzos de diciembre de ese año) a publicaciones con apenas trece «me gusta», dos «compartidos» y ningún comentario.

El caso de Cora Gamarnik se suma al de otras figuras públicas, como es el caso del periodista Sebastián Lacunza, quien en septiembre de 2023 publicó en su cuenta de X que se había percatado de que le habían aplicado un shadow banning, debido a que «las ‘impresiones’ bajaron a un quinto de lo habitual y se revirtieron abruptamente las interacciones». Además, aseguró que no aparecía en el resultado de búsqueda de X entre no seguidores, lo que implicaba un enorme impacto en cuanto al alcance y la visibilidad de su cuenta.

Emma y Yo: cuando reducen el alcance de la educación sexual integral en Instagram

«He estado a punto de tirar la toalla y decir, ya, a la mierda..., que quede el material ahí, qué lindo, ya se acabó, ya no quiero más».[46] La frustración en las palabras de Alesia Lund, creadora del proyecto de educación sexual @emmayyoperu[47] en Instagram, revela la realidad invisible que enfrentan educadores digitales en América Latina. Sin advertencias ni explicaciones, vio cómo su cuenta de Instagram, que había cultivado hasta casi sesenta y ocho mil seguidores, comenzó a desvanecerse sistemáticamente ante sus ojos, perdiendo más de diez mil seguidores en dos años mientras sus publicaciones caían de quince mil o veinte mil visualizaciones a apenas trescientas o quinientas. Lo que describe Alesia encuadra en lo que hemos descrito en esta investigación. Una práctica de moderación de contenidos que actúa como un fantasma en el algoritmo: no deja rastros, no envía notificaciones, simplemente hace desaparecer el contenido del radar público.

Emma y Yo nació en 2019 como una respuesta urgente a la escasez de contenido accesible sobre educación sexual en el contexto latinoamericano. A través de ilustraciones, infografías y material didáctico cuidadosamente diseñado, el proyecto se convirtió rápidamente en un referente regional para hablar de sexualidad con niños, niñas, adolescentes y jóvenes adultos. Su enfoque combina elementos informativos con una estética atractiva que desmitifica temas tabú —desde anatomía hasta consentimiento— utilizando un lenguaje directo y accesible. Con dos libros publicados y una comunidad mayoritariamente de mujeres jóvenes entre 18 y 24 años, la iniciativa de Alesia logró llenar un vacío educativo crítico en un continente donde la educación sexual institucional sigue siendo insuficiente. Su crecimiento orgánico en redes reflejaba no solo la calidad del contenido, sino también la enorme necesidad social de información confiable sobre un tema históricamente silenciado.

Lo más revelador del caso es la ausencia total de comunicación por parte de la plataforma. A diferencia de una suspensión tradicional, donde existe notificación por violación de normas, Alesia nunca recibió advertencia alguna sobre su contenido educativo sobre sexualidad. Esta invisibilización progresiva ha ocurrido paralelamente a cambios en las políticas de contenido de Meta, coincidiendo con el período pospandemia, cuando diversas cuentas educativas sobre sexualidad comenzaron a reportar problemas similares, cuenta Alesia. «Comencé a ver cuentas gringas que yo seguía que los estaban *banneando*, lo vi de lejos, y dije, esta vaina se nos viene ahorita». Tiempo después, Alesia comenzó a perder seguidores y sus publicaciones comenzaron a verse menos. El patrón se confirma al comparar con otras creadoras: contenidos sobre temas considerados controversiales (educación sexual, feminismo, derechos reproductivos) experimentan significativamente menor distribución que otros contenidos.

46 Entrevista realizada por Observacom y Digital Action (2025) a Alesia Lund, 14 de marzo de 2025.
47 Emma y Yo Perú. (14 de abril de 2025). Instagram. https://www.instagram.com/emmayyoperu/

El impacto de esta moderación invisible, en la sombra y opaca, trasciende lo numérico y alcanza dimensiones emocionales y profesionales significativas. «Mi producción ha bajado un montón», confiesa Alesia, describiendo ciclos de frustración que la llevaron a pausar su trabajo por semanas enteras. «¿Para qué miércoles hago el esfuerzo, si ni siquiera se ve lo que hago?», se preguntaba. Esta erosión gradual del alcance no solo afecta la motivación creativa, también impacta directamente la sostenibilidad de proyectos educativos independientes, orillando a creadores a migrar entre plataformas constantemente (de Facebook a Instagram, luego a TikTok y YouTube) en búsqueda de visibilidad. Es una situación, dice Alesia, que realmente desgasta. Por eso, está tratando de olvidarse de Instagram y enfocarse en plataformas como YouTube para ver si «al fin, por primera vez en mi vida puedo monetizar con el proyecto».

Este caso evidencia cómo el shadow banning opera como un mecanismo de control sin transparencia que afecta desproporcionadamente a voces que abordan temas considerados sensibles. El algoritmo parece penalizar términos como *sexualidad, vulva* o *educación sexual,* aun cuando se presentan en contextos educativos, por ello, Alesia dejó de usar algunos *hashtags*. Lo más preocupante es que, a diferencia de la censura tradicional, este sistema ofrece nula capacidad de apelación: sin recibir notificaciones sobre qué normas se han infringido, las personas creadoras no pueden ajustar su contenido para cumplir con lineamientos específicos. Esta forma de control invisible representa un desafío fundamental para la libertad de expresión en entornos digitales, particularmente para quienes abordan temas de salud sexual desde perspectivas educativas y feministas en América Latina.

El testimonio de Alesia apunta también a un fenómeno más amplio que afecta a las cuentas de educadores en toda la región y que está asociada al shadow banning: la comercialización de la atención en plataformas como Instagram está ejerciendo presión para monetizar el alcance. «Nos están orillando a que tengamos que pagar», señala, explicando cómo incluso cuentas nuevas sin contenido polémico enfrentan limitaciones severas si no invierten en promoción. «Antes el crecimiento era orgánico, rápido, y ahora el crecimiento es lentísimo, las vistas no se dan», cuenta Alesia. Esta realidad plantea interrogantes sobre el acceso democrático a la información, especialmente cuando contenidos educativos sobre sexualidad, considerados de interés público, quedan restringidos por algoritmos comerciales que no distinguen entre contenido sensible y contenido educativo, creando así un doble estándar particularmente perjudicial para iniciativas sin fines de lucro.

Finalmente, los patrones descritos por Alesia sugieren que existe una forma de discriminación algorítmica basada en temáticas. La priorización que tienen las plataformas sobre personas hablando sobre fotos e ilustraciones revela cómo las plataformas privilegian ciertos formatos sobre otros, lo que afecta desproporcionadamente a creadores que abordan temas sensibles mediante ilustraciones o textos educativos. Este sesgo técnico intensifica la invisibilización al mostrar, por un lado, que los videos de entretenimiento sin contenido educativo sobre sexualidad circulan libremente, y por el otro, que el material infográfico con un enfoque pedagógico científico queda relegado.

Cuando los algoritmos juzgan cuerpos: el caso Love&Lust contra la invisibilización digital

Esta discriminación algorítmica y shadow banning no solo afectan al contenido educativo en temas de sexualidad. Las empresarias independientes que construyen comunidades en torno a productos considerados «sensibles» enfrentan el mismo muro invisible de silenciamiento. El caso de Paula Labra, con su negocio de lencería inclusiva, también ofrece una muestra de este fenómeno.

«Grabé a una amiga buscando mi cuenta y simplemente no existía»,[48] relató en una entrevista. Con estas palabras Paula, propietaria del e-commerce de lencería (@lovelust.cl) con doscientos cuarenta mil seguidores, describe el momento en que documentó lo que miles de creadores sospechan pero pocas veces pueden probar: su cuenta había sido sometida a shadow banning, volviéndola prácticamente invisible en la plataforma que sustenta gran parte de su negocio.

El caso de Love&Lust expone con claridad los mecanismos opacos de moderación que afectan desproporcionadamente a los creadores de contenido. Paula documenta cómo Instagram aplica criterios discriminatorios basados en tipos de cuerpo. «En TikTok me pasaba que a mujeres flacas no las bajaban, a mujeres voluptuosas sí. Más gorditas y más [pecho], chao, me bajaban los posts». Este sesgo algorítmico evidencia una forma de censura que trasciende las políticas explícitas y reproduce prejuicios contra cuerpos no normativos, afectando particularmente el contenido que celebra la diversidad corporal femenina o muestra productos relacionados con la anatomía, como prótesis de pezones posmastectomía o ropa interior menstrual.

Los efectos del shadow banning son devastadores y cuantificables. «Lo sentí en las ventas. Era como que no existíamos, siendo una marca que podía estar invirtiendo tres mil dólares mensuales en publicidad», relata Paula. Este testimonio muestra como Meta cobra por servicios publicitarios mientras simultáneamente limita el alcance de las mismas cuentas que pagan por visibilidad. Un caso que, como veremos más adelante, también se extiende a otro tipo de servicios de la plataforma. La situación ha creado una nueva forma de ilusión algorítmica, en la que creadores pagan cuarenta y cinco dólares mensuales por Meta Verified «solo por miedo», sin garantía alguna de protección contra la invisibilización. «Estoy gastando más de quinientos dólares al año por miedo, eso es una estupidez», lamenta. El impacto psicoemocional es igualmente significativo: «Vivo con miedo, antes de quedarme dormida, siempre pido que no se me vaya la cuenta», una ansiedad constante que afecta su bienestar y capacidad creativa.

La respuesta de Paula ante esta censura sistemática revela la ingeniosa resistencia de creadores latinoamericanos frente a algoritmos opacos. Desarrolló estrategias comunitarias organizando

48 Entrevista realizada por Observacom y Digital Action (2025) a Paula Labra, 19 de marzo de 2025

«un grupo de amigas influencers» para que repostearan constantemente su contenido, buscando romper el shadow banning mediante interacciones masivas. Simultáneamente, documenta meticulosamente cada instancia de censura, guardando capturas de pantalla que evidencian el trato diferencial hacia cuentas pequeñas versus grandes marcas como Calvin Klein o Savage X Fenty, que pueden mostrar contenido mucho más explícito sin consecuencias. El caso de Paula es una muestra de una cuenta que, si bien no aborda asuntos de interés público, sí ilustra cómo la moderación algorítmica está creando un paisaje digital donde ciertos cuerpos y temas son sistemáticamente invisibilizados. Y obliga a pequeños emprendedores a destinar recursos considerables no a mejorar sus productos, sino a luchar ciegamente contra un sistema opaco que somete su existencia a un miedo constante de ser expelidos.

06.

EL SHADOW BANNING EN LAS REGLAS DE LAS PLATAFORMAS DIGITALES

Uno de los objetivos de esta investigación era indagar qué información brindan las plataformas digitales a sus usuarios sobre las medidas de moderación de contenido que afectan el alcance y la visibilidad de sus publicaciones o de sus cuentas —conocidas como shadow banning—, así como las posibilidades de apelar o impugnar estas decisiones.

En este apartado se examina qué dicen los términos y condiciones de X y de Meta (empresa matriz de las plataformas Facebook e Instagram, cuyas reglas se encuentran unificadas), así como también declaraciones de sus directivos y voceros, para identificar cómo estas plataformas definen, explican y regulan la reducción de alcance o reducción de visibilidad, y qué tipo de garantías y de derecho a defenderse ofrecen, si es que hay, a las personas usuarias afectadas.

Términos y condiciones de X

En el caso de las reglas de X, la plataforma incluye un apartado específico sobre las medidas que puede aplicar cuando considera que un contenido infringe sus normas. Allí distingue entre acciones de distintos niveles: medidas a nivel del post o la publicación, medidas a nivel de la cuenta o medidas en mensajes directos.

Entre las medidas que X puede aplicar sobre un post, figura expresamente la posibilidad de «limitar la visibilidad del post». En sus normas se explica que toman medidas en cuanto al contenido.

> Cuando hay un post específico que incumple las Reglas de X, incluidos aquellos en los que se comparten o reproducen otros posts, ya sea porque se publican capturas de pantalla, se citan posts o se comparten las URL de otros posts que incumplen las Reglas de X.

Las limitaciones de visibilidad del post la explican de la siguiente manera: «Cuando corresponda, restringiremos el alcance de los posts que incumplan nuestras políticas y generen una experiencia negativa para otros usuarios; para ello, haremos que el post sea más difícil de descubrir en X». Las medidas posibles incluyen:

1. Excluir el post de los resultados de búsqueda, las tendencias y las notificaciones de recomendaciones.
2. Eliminar el post de las cronologías «Para ti» y «Siguiendo».
3. Limitar la visibilidad del post al perfil del autor.
4. Posicionar el post más abajo en las respuestas.
5. Restringir los «me gusta», las respuestas y los reposts, así como la posibilidad de citar el post, guardarlo, compartirlo, anclarlo al perfil o editarlo.

A partir de abril de 2023, X comenzó a etiquetar públicamente los posts que han sido identificados como infractores de sus normas, informando tanto a los autores como a los lectores que la visibilidad del post ha sido limitada. Los autores tienen la posibilidad de solicitar una revisión de estas etiquetas si consideran que la limitación de visibilidad fue

aplicada incorrectamente. Sin embargo, no se especifica claramente cómo se debe realizar esta solicitud de revisión, a diferencia de los procedimientos detallados que existen para apelar suspensiones o bloqueos de cuentas.[49]

Además, X contempla una «excepción por interés público»[50] para ciertos contenidos que, a pesar de infringir las normas, se considera que deben permanecer accesibles debido a su relevancia pública. Esta excepción se aplica principalmente a publicaciones de cuentas de alto perfil que representan a miembros actuales o potenciales de organismos gubernamentales o legislativos. En estos casos, el post se coloca detrás de un aviso y se limita su visibilidad, pero permanece accesible en la plataforma.

Estos son los criterios para las excepciones:

1. El post incumple una o más de las reglas de X.
2. El post lo compartió una cuenta de alto perfil.
3. La cuenta representa a un miembro actual o potencial de un organismo gubernamental o legislativo local, estatal, nacional o supranacional:

    a. titulares actuales de un puesto de liderazgo elegido o designado en un organismo gubernamental o legislativo;
    b. candidatos o nominados para cargos políticos;
    c. partidos políticos registrados.

Términos y condiciones de Meta (Facebook e Instagram)

Meta, la empresa matriz de Facebook, Instagram, Threads y Messenger, incluye en sus documentos oficiales referencias explícitas a la práctica de reducir la visibilidad de ciertos contenidos, aun cuando estos no infrinjan sus normas comunitarias.[51] Esta medida se enmarca dentro de su política de «curación» de contenidos, estructurada desde 2016 en torno a tres ejes: suprimir, reducir e informar.

El pilar de «suprimir» posts es el más fácilmente reconocido y reconocible, ya que involucra las acciones de moderación de contenido clásicas como la remoción de contenidos y la eliminación de cuentas de usuarios. Por otro lado, según se explica en el apartado Reducir la distribución de contenido problemático,[52] en el Centro de Transparencia de la empresa, el enfoque de «reducción» busca limitar la circulación de lo que denominan «contenido problemático» que, si bien no viola directamente sus reglas, puede «generar experiencias negativas» o considerarse de «baja calidad». En estos casos, Meta declara que

49 X Centro de ayuda. Solicitar la revisión de una cuenta bloqueada o suspendida. https://help.x.com/es/forms/account-access/appeals/redirect
50 X Centro de ayuda. Acerca de las excepciones de interés público en X. https://help.x.com/es/rules-and-policies/public-interest
51 Desde el 12 de noviembre de 2024, Meta unificó sus normas comunitarias en una sola que aplica a sus cuatro servicios de redes sociales y mensajería: Facebook, Instagram, Messenger y Threads. Las normas comunitarias pueden leerse completas aquí: https://transparency.meta.com/es-es/policies/community-standards/
52 Meta. (25 de abril de 2025). Reducimos la distribución de contenido problemático. https://transparency.meta.com/es-es/enforcement/taking-action/lowering-distribution-of-problematic-content/

reduce su distribución en el *feed* y en las recomendaciones, sin eliminarlo ni notificar al usuario que ha sido penalizado.

La empresa tiene una clasificación amplia y ambigua de «contenido problemático» al que le pueden reducir la distribución:

- Contenido de baja calidad, como *clickbaits* y señuelos de interacción.
- Enlaces a sitios web que están repletos de anuncios, se cargan lentamente o no funcionan.
- Comentarios de baja calidad que se copian y pegan reiteradas veces.
- Contenido con originalidad limitada que se reutiliza principalmente de otras fuentes.
- Videos de baja calidad que usan indebidamente los formatos de video o video en directo.
- Información errónea y desinformación.
- Contenido de creadores que infringen reiteradamente nuestras políticas.

Como puede observarse, el listado incluye desde *clickbaits* y comentarios repetitivos hasta contenidos sin originalidad y publicaciones marcadas como desinformación. Es importante destacar que esta última categoría puede derivar en una penalización algorítmica sin que exista una evaluación humana ni un canal de apelación accesible para los usuarios afectados.

Meta justifica esta práctica como un mecanismo para priorizar la experiencia de los usuarios. Sin embargo, la empresa no ofrece notificaciones ni vías de apelación accesibles en estos casos, lo que impide que las personas usuarias comprendan si están siendo penalizadas y por qué. Esta falta de transparencia es particularmente relevante cuando las decisiones automatizadas afectan la circulación de contenido legítimo o vinculado a temas de interés público.

Además, las pautas de recomendación que Meta aplica para Facebook e Instagram refuerzan este tipo de prácticas. En ellas se establece que el contenido sugerido —como el que aparece en «Explorar», «Sugerencias para ti» o los reels— se rige por criterios internos destinados a evitar la amplificación de materiales que la plataforma considera inadecuados o irrelevantes para ciertos públicos. Sin embargo, no se explicitan los parámetros exactos que guían estas decisiones ni se ofrecen mecanismos transparentes para que las personas usuarias comprendan por qué su contenido ha dejado de circular con normalidad.

A diferencia de lo que ocurre con otros mecanismos de moderación de contenido —como la remoción de publicaciones o la suspensión de cuentas—, en el caso de la «reducción de distribución de contenido problemático», Meta no contempla en sus políticas la posibilidad de apelar ni solicitar revisión de la medida. Esto profundiza la opacidad del proceso, ya que las personas usuarias no reciben notificación alguna, ni tienen herramientas para revertir o discutir las decisiones que afectan directamente la visibilidad de sus publicaciones.

Entre las normas comunitarias de Meta también se incluyen restricciones específicas para el contenido vinculado al cannabis y sus derivados. En la sección sobre Bienes Regulados, la empresa establece que puede restringir publicaciones que «coordinen o promuevan (es decir, que hablen positivamente,

fomenten el uso o proporcionen instrucciones de uso o fabricación) la marihuana y productos que contengan THC o componentes psicoactivos relacionados».[53] Si bien la política no implica necesariamente la eliminación de las cuentas que compartan este tipo de contenidos, su inclusión en las normas permite a la empresa aplicar mecanismos de reducción de visibilidad o limitación del alcance, al menos entre usuarios menores de 18 años. Este marco regulatorio podría estar relacionado con el caso de la cuenta de Instagram Muypaola, cuya creadora denunció una caída abrupta en las interacciones y visibilidad de sus publicaciones. Sin embargo, en su caso, la limitación del alcance no parecería solo aplicarse a menores de 18 años, sino a usuarios en general.

53 Meta. Bienes y servicios restringidos. https://transparency.meta.com/es-la/policies/community-standards/regulated-goods/

07.

CONCLUSIONES

Esta investigación permitió identificar que el shadow banning es un conjunto de prácticas cada vez más frecuentes en las principales plataformas digitales, y que tiene un impacto significativo sobre la libertad de expresión, la circulación de información de interés público y la participación democrática en entornos digitales.

A través del análisis de casos concretos en América Latina, de entrevistas con expertos y del examen de los términos y condiciones de empresas como Meta (Facebook e Instagram) y X (antes Twitter), se evidenció que estas formas opacas de moderación operan como mecanismos de silenciamiento algorítmico que afectan a creadores de contenido, activistas, periodistas y comunidades históricamente marginadas. Ocultan contenidos mediante prácticas que los usuarios no pueden detectar.

Uno de los principales hallazgos es que la reducción de visibilidad puede tener consecuencias equiparables —e incluso más duraderas— que la remoción directa de contenido. Aunque los contenidos no son eliminados, su circulación se ve gravemente restringida, afectando tanto su alcance como la posibilidad de participación en el debate público. Esta penalización, que muchas veces actúa de forma automatizada, puede generar efectos desproporcionados que las personas usuarias solo advierten tras notar una caída repentina en las interacciones, visualizaciones o descubribilidad de sus cuentas.

Además, se constató que la transparencia de las plataformas respecto a estas prácticas es mínima o directamente inexistente. Meta reconoce que aplica medidas para «reducir la distribución de contenido problemático» —una categoría amplia y ambigua—, pero no informa a los usuarios cuándo estas medidas se activan ni ofrece vías claras de apelación. X, por su parte, menciona que puede limitar la visibilidad de ciertos contenidos y promete una posible revisión, pero no detalla los procedimientos ni garantiza su accesibilidad o efectividad. En la práctica, esto deja a las personas afectadas en una situación de indefensión, sin información suficiente para cuestionar la sanción ni herramientas para revertirla.

La investigación también evidencia que las decisiones sobre qué contenidos reducir en visibilidad no son neutrales y que los algoritmos de moderación no solo reproducen sesgos sociales existentes, también amplifican las desigualdades estructurales al limitar el acceso a voces disidentes o fuera de la norma dominante.

Por último, el estudio confirma que estas prácticas vulneran derechos individuales, así como también socavan el carácter público y democrático de los espacios digitales. Si las plataformas siguen operando sin transparencia, sin rendición de cuentas y sin mecanismos adecuados de defensa para las personas usuarias, el riesgo va más allá de la censura encubierta de ciertas voces: compromete el debate democrático en Internet.

Frente a este escenario, urge avanzar hacia marcos normativos y regulatorios que garanticen transparencia, debido proceso y derecho a defensa ante las medidas de reducción de alcance y visibilidad aplicadas por las grandes plataformas digitales.

OBSERVACOM

www. observacom.org