25 estándares para una internet más transparente en Colombia

UNIÓN EUROPEA

OBSERVACOM
Observatorio Latinoamericano de Regulación de Medios y Convergencia

unesco

EL VEINTE
JUSTICIA · LIBERTAD DE EXPRESIÓN

SOCIAL MEDIA 4PEACE

25 estándares para una Internet más transparente en Colombia

Los siguientes estándares sobre transparencia y rendición de cuentas de las grandes plataformas de redes sociales[1] son una iniciativa que forma parte del esfuerzo y el compromiso de numerosas organizaciones colombianas para avanzar en la protección del espacio público digital, recuperar una Internet más libre y abierta, lograr el pleno ejercicio de derechos fundamentales -entre otros, el de la libertad de expresión- en línea y contrarrestar el impacto de la desinformación, los discursos de odio y fortalecer la gobernanza de las grandes plataformas digitales.

Este esfuerzo parte de la evidencia de la falta de transparencia y rendición de cuentas de los procesos de moderación y curación de contenidos de unas pocas empresas transnacionales concentradas que se han convertido en *gatekeepers* de Internet que puede afectar a grupos en situación de vulnerabilidad como mujeres, personas afroamericanas, pueblos indígenas y la comunidad LGBTI, entre otros,

[1] Los siguientes estándares expresan una perspectiva asimétrica de la regulación, ya que están diseñados para alcanzar a las empresas dueñas de las "grandes plataformas" digitales comerciales que moderan contenidos de terceros y que se han convertido nuevos espacios públicos en línea de reconocida influencia en el debate público, en atención a la importancia y el impacto que sus decisiones empresariales tienen sobre el debate público y el intercambio de información, opiniones y bienes culturales, así como al ejercicio de la libertad de expresión y el debate público en nuestros países. No alcanza, por tanto, a plataformas pequeñas o sin fines de lucro, o servicios de mensajería instantánea.

además del trabajo de periodistas, medios independientes, comunicadores populares y personas defensoras de derechos humanos en general.

Y se asienta en el convencimiento de que hay que adoptar una actitud proactiva y alcanzar un consenso para garantizar la transparencia y la rendición de cuentas por parte de las plataformas digitales a través de obligaciones democráticas y asimétricas que complementen la adopción de buenas prácticas por parte de las propias empresas.

La falta de transparencia alimenta políticas públicas reactivas, desproporcionadas e innecesarias. Nuestras organizaciones seguirán rechazando estas propuestas para proteger esos derechos y una Internet libre y abierta, pero también creemos que una actitud proactiva respecto de la regulación puede ser un camino inteligente para limitar estos intentos y empoderar a nuestra gente. Las obligaciones de transparencia y rendición de cuentas pueden ser un primer paso, que debe darse tras un diálogo multisectorial amplio y abierto y el apoyo de organismos internacionales como la UNESCO y la Relatoría para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos (RELE-CIDH), para garantizar los derechos de las personas usuarias de las plataformas digitales, pero también para anticiparse a los intentos de regulación antidemocrática[2].

Esta propuesta de estándares de transparencia y rendición de cuentas de las grandes plataformas de redes sociales en Colombia fue elaborada mediante un análisis comparado de informes internacionales relevantes producidos por organismos internacionales y organizaciones con especialización en el tema, así como con fuentes regulatorias de alcance regional significativas.

Entre tales fuentes destacan la guía de "Directrices para la gobernanza de las plataformas tecnológicas"[3] publicada en 2023 por la UNESCO, el documento "Transparencia de la moderación privada de contenidos: una mirada de las propuestas de sociedad civil y legisladores de América Latina" -del cual OBSERVACOM es redactor-, la "Declaración latinoamericana sobre Transparencia de las plataformas de Internet" y "Dejar entrar el sol: transparencia y responsabilidad en la era digital" de la UNESCO.

También se consideraron como documentos claves los "Principios de Santa Clara 2.0 sobre la transparencia y la responsabilidad en la moderación de contenidos"[4], los "Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta" (OBSERVACOM, FLIP y otras organizaciones dela región y el informe sobre "Inclusión digital y gobernanza de contenidos en internet"[5] de la Relatoría Especial para la Libertad de Expresión de la CIDH.

El análisis buscó identificar principios y prácticas de regulación que sean compatibles con los más altos estándares de protección a los derechos humanos, con especial énfasis en los siguientes aspectos: alcance de la definición de transparencia, la

[2] Declaración Latinoamericana sobre Transparencia de las Plataformas de Internet, 24 de noviembre de 2021
[3] https://unesdoc.unesco.org/ark:/48223/pf0000387360/PDF/387360spa.pdf.multi
[4] https://santaclaraprinciples.org/es/
[5] https://www.oas.org/es/cidh/expresion/informes/Inclusion_digital_esp.pdf

transparencia y asequibilidad de los términos de servicio, obligaciones asimétricas de transparencia, obligaciones de rendición de cuentas, reglas de transparencia algorítmica, notificaciones frente a remociones de contenido o suspensión de cuentas, así como la transparencia en aspectos comerciales relevantes (como el pautado de anuncios políticos en periodos electorales), entre otros.

Por otra parte, la elaboración de la propuesta de OBSERVACOM y El Veinte implicó un trabajo de construcción de acuerdos progresivos en torno a esos estándares de transparencia, que implicaron una serie de consultas, reuniones virtuales y talleres presenciales durante todo el proceso que involucró a las organizaciones y entidades que participaron del proyecto SM4P en Colombia. Se buscó -en contacto con la coordinación del proyecto y organizaciones aliadas como Articulo 19 y el ICESI- que este proceso se entrelazar con la construcción de una alianza de organizaciones de sociedad civil para la incidencia en estos temas, con el objetivo de que este documento final pueda ser apropiado por las organizaciones que la integren, de forma que integre sus iniciativas y su plan de acción.

Transparencia de las reglas de juego

1. Los términos de servicio (TdS) de todas las plataformas de contenidos, así como otros documentos complementarios (como guías o directrices de aplicación de contenidos) deberían estar redactados de manera clara, precisa, inteligible y accesible para todos los usuarios en nuestro idioma nacional, es decir, español.[6] En el caso de Colombia en necesario observar que en suma de lo anterior, para garantizar la accesibilidad de estos documentos se debería contar con versiones en lenguas indígenas y accesibles para personas con discapacidades sensoriales.[7]

2. Cualquier cambio de los términos de servicio y normas comunitarias se debería informar claramente al usuario, con su justificación y solicitud de consentimiento explícito, con un tiempo de aviso razonable[8].

3. Informar a sus usuarios, en un lugar central y de fácil acceso, qué tipos de contenido y actividades están prohibidos y pueden llevar a la eliminación,

[6] Expresado en la Guía de Directrices para la gobernanza de las plataformas digitales de la UNESCO de esta manera: “Las plataformas deben tener sus condiciones de servicio completas disponibles en el idioma oficial y en los idiomas principales de cada país donde operan, garantizar que puedan responder a las personas usuarias en su propio idioma y procesar sus reclamaciones por igual, así como tener la capacidad de moderar y seleccionar el contenido en el idioma de la persona usuaria. Los traductores automáticos de idiomas pueden utilizarse para ofrecer una mayor accesibilidad lingüística, pero debe controlarse su fidelidad debido a sus limitaciones técnicas”. (párr. 119)

[7] En seguimiento a la recomendación de la Relatoría Especial para la Libertad de Expresión de la CIDH en su informe Inclusión digital y gobernanza de contenidos en internet: “Disponer la información en todos los idiomas, incluyendo lenguas indígenas, y hacer accesible dicha información a personas con discapacidad, utilizando lenguaje claro y evitando terminología técnica” (párr. 295)

[8] Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta (párr. 2.3)

desindexación o reducción significativa del alcance de sus expresiones, o a la suspensión o bloqueo de su cuenta, sea de forma permanente o temporal[9].

4. Información sobre el número de moderadores humanos empleados o subcontratados que tratan sobre asuntos locales, si se encuentran en Colombia o cuál es su ubicación geográfica y la naturaleza de su experiencia y conocimiento de nuestro idioma y de nuestro contexto local, y si reciben capacitación en temas de libertad de expresión y otros derechos humanos[10].

5. En qué casos, cuándo y cómo aplica la automatización de remoción de contenidos y cuándo y cómo aplica la revisión humana de contenidos[11]. Esto debería incluir información sobre el grado en que la moderación humana es tercerizada a otras empresas.

6. Información sobre qué criterios se utilizarán para adoptar decisiones teniendo en cuenta el contexto, la amplia variación de matices idiomáticos y el significado y las particularidades lingüísticas y culturales de los contenidos sujetos a una posible restricción[12].

Transparencia algorítmica

7. Dar a conocer los criterios utilizados por algoritmos para la moderación y curación (ordenamiento, priorización, reducción de alcance, amplificación, recomendación o direccionamiento) de contenidos, explicando su impacto en la visibilidad de estos y diferenciando entre los controlables por los usuarios y los que no lo son, explicitando los efectos para el usuario[13].

8. Informar sobre cuándo y cómo se utilizan los procesos automatizados sobre la moderación y curación de contenidos, los criterios clave utilizados por los procesos automatizados para tomar decisiones, y cuáles son las categorías y tipos de contenidos en los que se utilizan estos procesos[14].

9. Informar cuál es el grado de supervisión humana de los procesos automatizados, incluida la posibilidad de que los usuarios soliciten la revisión humana de cualquier decisión de moderación de contenidos automatizada[15].

---

[9] Las referencias de este punto son los Principios de Santa Clara 2.0 sobre la transparencia y la responsabilidad en la moderación de contenidos (principio "Normas y políticas comprensibles") y los Estándares para una regulación democrática de las grandes plataformas (párr. 2.2)
[10] Directrices para la gobernanza de las plataformas digitales (párr. 115 i)
[11] Estándares para una regulación democrática de las grandes plataformas (párr. 3.4)
[12] Ídem anterior
[13] Inclusión digital y gobernanza de contenidos en internet (párr. 295 c)
[14] Principios de Santa Clara 2.0. Por otra parte, es fundamental que esta información no se limite e la mera enunciación de técnicas genéricas como "análisis de sentimiento (*sentiment analysis*)" o "visión por computador (*computer vision*)", "redes neuronales (*neural networks*)", "aprendizaje profundo (*deep learning*)" si no que debe dar a conocer de manera concreta el tipo de tecnología implementada.
[15] Ídem anterior

10. En las acciones de priorización o recomendación de contenidos en línea accesibles al usuario debería ser claramente identificable la naturaleza comercial de las comunicaciones, el contenido patrocinado, así como la propaganda electoral o política, de forma clara, identificando al contratante y sin generar dudas acerca de su significado y siendo transparente sobre los metadatos del contenido (precios, etc.)[16].

11. Informar sobre la manera en que los datos personales de las personas usuarias se recogen, utilizan, revelan, conservan y difunden y el tratamiento que se les da, incluido qué datos personales y sensibles se utilizan para tomar decisiones algorítmicas con fines de moderación y curación de contenido. Esto también incluye cómo se comparten los datos personales con otras entidades y qué datos personales obtiene la plataforma indirectamente, por ejemplo, a través de la elaboración de perfiles de usuario o la interoperabilidad con otras partes del ecosistema digital[17].

12. Las plataformas digitales deben facilitar a universidades y personas investigadoras autorizadas el acceso a los datos no personales y a los datos seudónimos necesarios para comprender el impacto de las plataformas digitales[18].

Empoderamiento de las personas usuarias

13. Las empresas deben notificar de manera oportuna y efectiva a cada usuario cuyo contenido sea eliminado, cuya cuenta sea suspendida, o cuando se tome alguna otra medida debido al incumplimiento de las normas y políticas del servicio, sobre el motivo de la eliminación, suspensión o acción. Cualquier excepción a esta regla, por ejemplo, cuando el contenido sea *spam*, *phishing* o *malware*, debe estar claramente establecida en las normas y políticas de la empresa[19].

Las notificaciones deberían:

- Incluir la cláusula específica de las normas comunitarias o la ley que se supone ha violado el usuario y ser lo suficientemente detallada para permitir al usuario identificar específicamente el contenido restringido[20].
- Incluir información sobre cómo se detectó, evaluó y eliminó o restringió el contenido o cuenta.
- Estar disponibles en una forma duradera que sea accesible incluso si la cuenta del usuario es suspendida o cancelada.

[16] Estándares para una regulación democrática de las grandes plataformas (párr. 3.3)
[17] Directrices para la gobernanza de las plataformas digitales (párr. 115 j)
[18] Directrices para la gobernanza de las plataformas digitales (párr. 116)
[19] Principios de Santa Clara 2.0 (“Avisos”)
[20] Estándares para una regulación democrática de las grandes plataformas (párr. 3.5)

- Estar en el idioma de la publicación original o en el idioma de la interfaz de usuario seleccionado por el usuario.
- Proporcionar a los usuarios información sobre los canales de asistencia disponibles y cómo acceder a ellos[21].

14. Las empresas deben informar de manera clara, de fácil acceso y en el idioma utilizado por el usuario, cuáles son los procesos de apelación ante decisiones que afecten contenidos de sus usuarios. Las empresas deben ofrecer una oportunidad significativa de apelar a tiempo las decisiones de eliminar contenidos, mantener contenidos que hayan sido marcados, suspender una cuenta o tomar cualquier otro tipo de acción que afecte a los derechos humanos de los usuarios, incluido el derecho a la libertad de expresión[22].

15. Las empresas deben garantizar que la apelación incluya

- Un proceso que sea claro y fácilmente accesible para los usuarios, con detalles de la línea de tiempo proporcionada a los que los utilizan, y la capacidad de seguir su progreso.
- Una revisión humana por parte de una persona o panel de personas que no hayan participado en la decisión inicial.
- Que la persona o el grupo de personas que participen en la revisión estén familiarizados con el lenguaje y el contexto cultural del contenido relevante para el recurso.
- Una oportunidad para que los usuarios presenten información adicional en apoyo de su recurso que se tendrá en cuenta en la revisión.
- Notificación de los resultados de la revisión y una exposición de motivos suficiente para que el usuario pueda entender la decisión[23].

Transparencia electoral

16. Información sobre anuncios políticos, en especial durante periodos electorales, incluido el importe, el autor y quienes financian los anuncios, el alcance pautado y concretado, debe conservarse en una biblioteca de acceso público en línea y ser visible por parte del receptor del mensaje[24].

Esta información debería incluir:

- Contenido anunciado.
- Importes invertidos.
- Responsable del pago.
- Periodo de impulso de las publicaciones.
- Características de los grupos de población que componen la audiencia de la publicidad contratada.

[21] Principios de Santa Clara 2.0 (“Avisos”)
[22] Principios de Santa Clara 2.0 (“Apelación”)
[23] Ídem anterior
[24] Directrices para la gobernanza de las plataformas digitales (párr. 115 l)

- Cantidad de personas afectadas por anuncios.

17. En el marco de periodos electorales deberían informar qué tipos de contenidos están prohibidos por la empresa y serán eliminados en relación con desinformación, discurso de odio e integridad electoral, entre otros, con orientaciones detalladas y ejemplos de contenidos permitidos y no permitidos; los tipos de contenidos contra los que la empresa tomará medidas distintas de la eliminación, como la reducción de la clasificación algorítmica, con orientaciones detalladas y ejemplos sobre cada tipo de contenido y acción; y las circunstancias en las que la empresa suspenderá la cuenta de un usuario, ya sea de forma permanente o temporal[25].

18. Informar sobre prácticas de publicidad y recopilación y uso de datos personales para targetizar la publicidad y resultados de la evaluación de impacto de los sistemas publicitarios en los derechos humanos y la igualdad de género[26].

19. Los contenidos generados exclusivamente por máquinas deben etiquetarse como tales (*deep fakes* y similares)[27].

Rendición de cuentas

20. Publicar reportes periódicos con información específica y desagregada[28] sobre las restricciones de contenido adoptadas en el país, tales como remociones, reducción de alcance, suspensión y bloqueo de cuentas, incluyendo acciones ante peticiones gubernamentales, órdenes de tribunales, requerimientos de privados, y como resultado de sus propias normas y políticas[29].

21. Publicar informes sobre eventos de alto impacto público, como protestas masivas, alteraciones al orden público, situaciones de conflictividad social,

[25] Principios de Santa Clara 2.0 (“Normas y políticas comprensibles”)
[26] Directrices para la gobernanza de las plataformas digitales (párr. 115 m)
[27] Directrices para la gobernanza de las plataformas digitales (párr. 115 o)
[28] Los Principios de Santa Clara desarrollan así este punto (“Números”):
- Número total de contenidos intervenidos y cuentas suspendidas.
- Número de apelaciones a las decisiones de actuar contenidos o suspender cuentas.
- Número (o porcentaje) de apelaciones exitosas que resultaron en piezas de contenido o cuentas que se restablecieron, y el número (o porcentaje) de apelaciones infructuosas y;
- Número (o porcentaje) de apelaciones exitosas o infructuosas de contenidos marcados inicialmente por la detección automática.
- Número de publicaciones o cuentas restablecidas por la empresa de forma proactiva, sin ningún tipo de apelación, tras reconocer que habían sido actuadas o suspendidas erróneamente.
- Números que reflejen la aplicación de las políticas de incitación al odio, por grupo o característica objetivo, cuando sea evidente, aunque las empresas no deberían recopilar datos sobre grupos objetivo para este fin
- Números relacionados con la retirada de contenidos y las restricciones realizadas durante periodos de crisis, como durante la pandemia del COVID-19 y los periodos de conflicto violento.

[29] Inclusión digital y gobernanza de contenidos en internet (párr. 295 h) y Estándares para una regulación democrática de las grandes plataformas (párr. 6.1)

elecciones, entre otros. Éstos deben detallar si han existido contactos con entidades estatales, los mecanismos de moderación específicos implementados, y el impacto de tales mecanismos[30].

22. Informar sobre las demandas y solicitudes de los agentes estatales (incluidos los organismos gubernamentales, las autoridades reguladoras, los organismos encargados de hacer cumplir la ley y los tribunales)[31] para la eliminación de contenidos o la suspensión de cuentas, así como los motivos esgrimidos para ello, y cuál fue el resultado de la solicitud y sus razones[32].

23. Información relevante respecto a apelaciones sobre la eliminación, el bloqueo o la negativa de bloquear el contenido y el modo en que las personas usuarias pueden acceder al proceso de reclamación. Estos datos deben incluir información cuantitativa y cualitativa de las reclamaciones recibidas, gestionadas, aceptadas y rechazadas, así como de los resultados de tales reclamaciones, e información sobre las reclamaciones recibidas por los funcionarios estatales y las medidas adoptadas[33].

Transparencia estatal

24. Los Estados deben informar por sí mismos de su participación en las decisiones de moderación de contenidos, incluyendo datos sobre las demandas o solicitudes para que se actúe sobre los contenidos o se suspenda una cuenta, desglosados por la base legal de la solicitud[34], con informaciones sobre:

- Cuántos pedidos de datos de usuarios fueron realizados.
- Qué motivos se argumentaron para justificar tales pedidos.
- En qué marcos legales se apoyan las solicitaciones realizadas.
- Cuántos pedidos de remoción de contenidos o suspensión o bloqueo de cuentas fueron realizados.
- Cuál fue la respuesta de las empresas en cada caso[35].

---

[30] Inclusión digital y gobernanza de contenidos en internet (párr. 295 d)
[31] Los Principios de Santa Clara desarrollan así este punto ("Números"):
- El número de demandas o solicitudes realizadas por actores estatales para que se actúe sobre el contenido o las cuentas
- La identidad del agente estatal para cada solicitud
- Si el contenido fue marcado por una orden judicial/juez u otro tipo de agente estatal
- El número de demandas o solicitudes realizadas por los agentes estatales que fueron objeto de acción y el número de demandas o solicitudes que no dieron lugar a una acción.
- Si el fundamento de cada señalamiento fue una supuesta infracción de las normas y políticas de la empresa (y, en caso afirmativo, qué normas o políticas) o de la legislación local (y, en caso afirmativo, qué disposiciones de la legislación local), o ambas cosas.
- Si las acciones tomadas contra el contenido se basaron en una violación de las normas y políticas de la empresa o en una violación de la legislación local. (Números)

[32] Principios de Santa Clara 2.0 ("Participación del Estado en la moderación de contenidos")
[33] Directrices para la gobernanza de las plataformas digitales (párr. 115 k)
[34] Principios de Santa Clara 2.0 ("Promover la transparencia gubernamental")
[35] Estándares para una regulación democrática de las grandes plataformas (párr. 3.5)

25. Los Estados deben garantizar que no se prohíba a las empresas publicar información que detalle las solicitudes o demandas de retirada o aplicación de contenidos o cuentas que provengan de actores estatales, salvo cuando dicha prohibición tenga una base legal clara, y sea un medio necesario y proporcionado para lograr un objetivo legítimo[36].

[36] Principios de Santa Clara 2.0 (“Eliminar los obstáculos a la transparencia de las empresas”)

La presente publicación fue elaborada en el marco del proyecto Social Media 4 Peace de UNESCO y contó con el apoyo de la Unión Europea. Su contenido es responsabilidad exclusiva de El Veinte y Observacom y no necesariamente refleja los puntos de vista de la UNESCO o de la Unión Europea.