Contribución de Derechos Digitales sobre el rol de actores privados en el abordaje de los discursos no protegidos en el SIDH

Consulta de la RELE/ CIDH sobre discursos no protegidos, punto n. 13 sobre "Criterios para el abordaje de los discursos no protegidos en Internet y el rol de agentes privados en el marco jurídico interamericano."[1]

Sobre Derechos Digitales

Derechos Digitales[2] es una organización regional latinoamericana sin fines de lucro fundada en 2005, que se dedica a la defensa y promoción de derechos humanos en el entorno digital de modo a contribuir con sociedades más justas, inclusivas e igualitarias. Sus acciones combinan investigación, incidencia en políticas públicas y privadas, análisis de tecnologías, campaña y formación en derechos digitales y seguridad digital.

Estándar vigente sobre actores privados de internet

El papel de los actores privados de internet en la moderación o curaduría de discursos no protegidos, a la fecha no ha sido abordado por la Relatoría Especial para la Libertad de Expresión, RELE como tal, sino a partir de la moderación de contenidos aplicada a la curaduría de la expresión en línea. Respecto a esa labor, la RELE ha observado en distintas ocasiones[i] que los intermediarios de internet tienen, tres subtipos de obligaciones que pueden ser caracterizadas de una manera tripartita así:

Primero. Las obligaciones aplicables a la moderación como proceso:

- Desarrollar mecanismos de supervisión transparente y garantías independientes para abordar sus reglas privadas de moderación de contenidos que puedan ser contrarias al derecho internacional.
- Observar garantías mínimas de debido proceso, incluida la notificación oportuna a los usuarios cuando su contenido pueda ser objeto de una medida de sanción por la plataforma o servicio, y brindar al usuario la posibilidad de cuestionar dicha acción "ateniéndose exclusivamente a restricciones prácticas que sean líticas o razonables efectuado un control minucioso de las pretensiones planteadas"[ii]

Segundo. Las obligaciones aplicables a las políticas de moderación:

- Si desarrollan normas de moderación que excedan lo exigido legalmente, sus políticas deben ser claras y preestablecidas, estar basadas en criterios objetivamente justificables, no deben atender fines ideológicos o políticos, y ser fruto de la consulta con sus usuarios.
- Asegurar medidas para que sus políticas sean fácilmente consultables y comprensibles por sus usuarios, incluida información sobre cómo se aplican.

---

[1] Elaborado por Lucía Camacho y Michel Souza. Para consultas, comunicarse a lucia.camacho@derechosdigitales.org y michel.souza@derechosdigitales.org

[2] https://www.derechosdigitales.org/

- Prever en sus "términos y servicios y reglas de comunidad, [que las empresas] no deben limitar o restringir la libertad de expresión de manera desproporcionada o innecesaria".
- No obstaculizar el acceso a contenidos sobre elecciones y la disponibilidad de diversos puntos de vista para los usuarios.

Tercero. Las obligaciones fruto del impacto de las prácticas de moderación:

- Adoptar medidas de transparencia sobre el uso e impacto que pueden ocasionar herramientas automatizadas de moderación de contenidos, incluyendo información sobre en qué medida afectan, entre otros, la clasificación o eliminación de los mismos.
- Deben establecer "sistemas eficaces de vigilancia, evaluaciones de impacto, sistemas de denuncias accesibles y eficaces a fin de identificar los daños reales o potenciales a los derechos humanos causados por sus servicios o actividades".
- Deben disponer de mecanismos adecuados para proporcionar remedios apropiados a los afectados.
- Deben ajustar sus actividades y sistemas para prevenir futuros abusos a los derechos de sus usuarios.

En estos comentarios retomaremos ese trabajo adelantado por la RELE para enfatizar la importancia de construir criterios operativos que sirvan para transparentar y aumentar la rendición de cuentas de los mecanismos y políticas aplicadas a la moderación de contenidos, en especial de los contenidos no protegidos que llevan a cabo los actores privados de internet.

Moderación, discurso de odio y el estándar interamericano

En la actualidad, la moderación de contenidos a cargo de las plataformas de internet (en especial, los intermediarios que facilitan la transmisión de contenido generado por sus usuarios y que sirven de medio para la conexión entre éstos) es el mecanismo operativo principal y de mayor alcance a través del cual los discursos protegidos y no protegidos, son curados en línea.

Dicha moderación está sujeta a las políticas de cada plataforma que, a manera de silo o de jardín vallado, fija criterios propios sobre qué contenidos se consideran formas prohibidas del discurso que ameritan su remoción de la plataforma, la suspensión o la cancelación de la cuenta de usuario que los difunde –entre otras medidas que varían según la plataforma en cuestión-.

En sus políticas, las plataformas de internet suelen coincidir en la prohibición de cierto tipo de *discursos ilegales*. Entre estos se encuentran la prohibición de la difusión de contenidos sobre abuso sexual infantil; contenidos que promuevan formas extremas de violencia contra grupos protegidos (en razón a su raza, color, religión, idioma, origen nacional, etc.); y contenidos que promueven el terrorismo o el genocidio. Estos contenidos son considerados ilegales por diversos tratados internacionales.[iii]

El **discurso de odio**[iv], por su parte, es considerado ilegal[v] por el *corpus interamericano* cuando incentiva a la violencia contra una persona o grupo de personas en razón a características protegidas como lo son la raza, el color, la religión, la orientación sexual,

identidad de género o diversidad corporal[vi]. Pero cuando las expresiones o discursos de odio no supera el umbral de *incitar a la violencia*, aun cuando puedan denigrar, estigmatizar o discriminar a las personas en razón a ciertas categorías protegidas "*pueden ser* sometidas a la imposición de sanciones ulteriores de naturaleza civil o administrativa, o a recursos como el derecho a la rectificación y réplica"[vii], pues es un tipo de discurso que precisa ser expresado en tanto que es de interés que sea discutido, rebatido y puesto en duda con más discurso y contradiscurso.

En este último caso, tanto la CIDH como la RELE insistieron en su informe sobre Violencia por Prejuicio contra personas LGBTI que "las sanciones [ulteriores] no pueden estar dirigidas a inhibir o restringir la diseminación de ideas o información [porque] la prohibición jurídica de este tipo de discurso no eliminará el estigma, el prejuicio o el odio profundamente arraigados en las sociedades de América [por lo que] son necesarias otras acciones por parte del Estado, los medios y la sociedad en general"[viii].

Se trata de una linea divisoria en todo caso compleja y que se atenúa cuando "la línea que separa el discurso de odio del discurso incómodo y minoritario (…) [se refiere a las] expresiones que se dirigen a figuras públicas o líderes políticos"[ix].

Aquí, emergen retos en torno a la alineación de las plataformas con el estándar interamericano para que sus políticas integren esta distinción que da un tratamiento diferente al discurso de odio en razón a los resultados —o el daño— que genera. Alineación que, frente a otros estándares de derechos humanos a nivel internacional, no ex mayoritariamente expresa o manifiesta.

Solo por poner un ejemplo sobre los retos de alineación o compatibilidad con estándares en la materia, en 2019 el Relator para la Libertad de Expresión de las Naciones Unidas invitó[x] en su momento a las plataformas de internet a adherir y adoptar voluntariamente el Plan de Acción de Rabat, con el fin de que facilitase a dichas empresas la tarea de distinguir "cuándo un contenido específicamente definido -los mensajes o las palabras o las imágenes que componente el mensaje- merece una restricción". Para 2021 solo la Junta Asesora de Meta —mas no la compañía— había manifestado expresamente adherir a dicho estándar[xi].

La moderación y sus viejos problemas

Pero, además, emergen otros retos asociados a la manera en que las políticas que puedan integrar ese matiz, sean posteriormente aplicadas pues la moderación de contenidos como proceso —automatizado o manual— adolece de serias falencias.

La moderación de contenidos a escala es, en su propia naturaleza, imperfecta[xii]. En primer lugar ha probado ser operativamente *inconsistente*, bien porque ciertas formas de discurso ilegal no son detectados y pese a ello circulan en línea, o bien porque contenidos que no son ilegales, pero que son considerados como dañinos por las plataformas, son censurados por éstas. En segundo lugar, porque es sabido que para cierto tipo de plataformas, como las redes sociales, el discurso de odio —en general— genera mayor interacción, atención y enganche entre los usuarios, por lo que hay un incentivo detrás en

permitir su circulación aún cuando pueda ser contrario a las políticas de moderación de la propia plataforma.

Esta manera de operar, según un informe publicado en 2020 por Derechos Digitales, genera que las plataformas de internet en la práctica sean 'significativamente más restrictiv[a]s en materia de libertad de expresión que los estándares bajo los cuales, al menos teóricamente, deben regirse los Estados'[xiii].

Estos problemas se encuentran aunados, además, a la profunda *heterogeneidad de las políticas* de moderación incluso entre plataformas que hacen parte de un mismo ecosistema, como las redes sociales. Por ejemplo, para 2021, solo Facebook, pero no en Instagram, consideraba relevante el contexto a la hora de determinar la circulación o no de contenidos considerados como discurso de odio —ambas de propiedad de Meta-, mientras tanto solo Instagram permitía de manera excepcional su circulación en los casos en que el discurso de odio fuese socialmente relevante.[xiv]

Para entonces, en Twitter así como en TikTok no se reconocían excepciones sobre la relevancia social del discurso de odio para permitir su circulación. Sobre las categorías protegidas en la difusión de discursos de odio, tampoco había acuerdos. Por ejemplo, Facebook consideraba la casta como una categoría protegida, mientras que Instagram no. Twitter, por su parte, no consideraba el sexo como un atributo protegido, que sí lo era para Facebook, Instagram así como para TikTok.[xv]

Además, son políticas que incluso que pueden mutar sucesivamente en períodos muy cortos de tiempo. Por ejemplo, la política de Meta sobre discurso de odio cambió 4 veces en 2021, y hasta 2024 acumulaba casi una docena de modificaciones que, aunque sutiles, tienen un impacto relevante para sus usuarios.[xvi]

Este panorama se encuentra aunado también a la manera inconsistente con que una sola red social puede aplicar de manera diversa una misma política en casos que son análogos. La Junta de Supervisión de Meta ha señalado cómo el contexto –relevante en la política sobre discurso de odio aplicable en Facebook y que la Junta interpreta en razón a su adopción voluntaria del Plan de Acción de Rabat— no es por defecto un elemento considerado en la moderación automatizada, y solo se considera de manera excepcional en la moderación humana que interviene en los procesos de apelación de los contenidos supuestamente infractores.[xvii]

La falta de *transparencia* también es un elemento clave en esta ecuación de moderación imperfecta. En el caso de Meta, la propia Junta de Supervisión apunta a la ausencia ha puesto de presente que Meta falla en informar sobre cómo es que en la práctica busca apegarse a sus políticas de moderación de contenido en general.[xviii] Ausencia de transparencia que se agrava con el desmantelamiento reciente de la herramienta CrowdTangle que permitía a investigadores y periodistas analizar de manera independiente los contenidos que circulaban en Facebook e Instagram para analizar, entre otros, el fenómeno del discurso de odio, sus dinámicas, audiencias, impacto, actores, etc.[xix]

Recapitulando, los problemas descritos más arriba pueden ser caracterizados en (i) aquellos de *tipo sustantivo*, que giran en torno a la definición de los alcances y límites del discurso de odio en sintonía con los estándares interamericanos, y (ii) los de *tipo operativo*, en torno a los estándares sobre los procedimientos aplicables a la moderación para dotarlo de mayor transparencia y rendición de cuentas en tanto que proceso altamente falible.

Criterios para actores privados de internet

En el pasado hemos expresado sobre esta misma materia que "sería exigible, como estándar mínimo, que las normas relativas al contenido permitido en redes sociales y plataformas de internet se ajusten a los estándares globales en materia de derechos humanos, no solo en lo sustantivo que respecta al contenido permisible y prohibido, sino en lo referido a estándares de transparencia, garantías mínimas de proceso, apelación e información al usuario"[xx].

Creemos que la CIDH y la RELE pueden extender estos estándares a las empresas de internet en razón a la aplicabilidad de los Principios sobre Empresas y Derechos Humanos de la ONU, así como los estándares interamericanos sobre Empresas y Derechos Humanos elaborados por la REDESCA.

Si bien el papel de las empresas tecnológicas no ha sido explorado en los principios elaborados por la REDESCA —aunque eso último sí haga parte del más reciente plan de trabajo 2024-2026 de esa Relatoría, lo que debería impulsar acciones de cooperación conjunta entre la RELE y la REDESCA para ampliar el *corpus iuris* latinoamericano aplicable a empresas de internet—, sí que hay apreciaciones sobre la aplicación extraterritorial de la ley así como de rendiciones de cuentas aplicables a grandes empresas transnacionales que también son aplicables por analogía al caso de las empresas de Silicon Valley.

Por eso tomaremos aportes de los Principios de Santa Clara[xxi], propuesto por organizaciones de la sociedad civil, así como la publicación "Moderación de Contenidos desde una Perspectiva Latinoamericana"[xxii] elaborada por AlSur. La propuesta de criterios la elaboramos con base en la construcción de estándares que ha sido avanzada hasta ahora por la RELE y la CIDH:

Primero. Las obligaciones aplicables a la moderación como proceso:

- Desarrollar mecanismos de supervisión transparente y garantías independientes para abordar sus reglas privadas de moderación de contenidos que puedan ser contrarias al derecho internacional.
    - **Sugerimos:** Exigir a las plataformas de internet que despliegan procesos de moderación que publiquen informes estadísticos desagregados, estructurados y en formatos abiertos que permita evaluar a usuarios, academia, sociedad civil e investigadores el funcionamiento y diseño de los mecanismos de moderación, especialmente cuando son aplicados a la moderación de discursos de odio.

    - **Sugerimos:** Exigir a las plataformas de internet que despliegan procesos de moderación que sean culturalmente sensibles del lenguaje, dialectos y contextos sociales y políticos en que los que el discurso se inserta.
    - **Sugerimos:** Exigir a las plataformas de internet explicar la causal por la cual se motiva una decisión de moderación, el método de detección empleado –manual o automatizado—, y en el caso de remociones motivadas en decisiones judiciales, la explicación del fundamento legal o la autoridad que elevó dicha solicitud.
    - **Sugerimos:** Exigir a las plataformas de internet dotar de mayores recursos, humanos y financieros, a las áreas encargadas de la moderación humana de contenidos, así como de diseño de políticas seguras para las personas usuarias de las mismas.
- Observar garantías mínimas de debido proceso, incluida la notificación oportuna a los usuarios cuando su contenido pueda ser objeto de una medida de sanción por la plataforma o servicio, y brindar al usuario la posibilidad de cuestionar dicha acción "ateniéndose exclusivamente a restricciones prácticas que sean líticas o razonables efectuado un control minucioso de las pretensiones planteadas".[xxiii]
    - **Sugerimos:** Exigir a las plataformas de internet informar claramente sobre qué procesos tienen a la mano para solicitar la revisión de un discurso identificado por la plataforma de manera errónea o que circula pese a su denuncia. Facilitar el seguimiento y consulta de dicha queja, así como informar de los procesos de apelación y sus instancias y tiempos.

Segundo. Las obligaciones aplicables a las políticas de moderación:

- Si desarrollan normas de moderación que excedan lo exigido legalmente, sus políticas deben ser claras y preestablecidas, estar basadas en criterios objetivamente justificables, no deben atender fines ideológicos o políticos, y ser fruto de la consulta con sus usuarios.
    - **Sugerimos:** Describir claramente los tipos de mecanismos de moderación aplicados a los discursos de odio según el estándar interamericano (ejemplo: si se limitará su visibilidad, si se sancionará o suspenderá al usuario de la cuenta, o se retirará el contenido, etc.), así como su extensión en el tiempo.
    - **Sugerimos:** Exigir a las plataformas la creación de políticas sobre discurso de odio diferenciadas para figuras públicas, personajes electos o personas que ocupan cargos políticos; así como políticas sobre discurso de odio en contextos diferenciados como el electoral, por su alto impacto para la preservación de los valores democráticos en la región.
    - **Sugerimos:** Exigir a las plataformas la creación de mecanismos de múltiples partes para su consulta a propósito del diseño de políticas de moderación.
- Asegurar medidas para que sus políticas sean fácilmente consultables y comprensibles por sus usuarios, incluida información sobre cómo se aplican.
    - **Sugerimos:** Que las políticas de la comunidad o los términos de uso del servicio sean fácilmente consultables por las personas, y que su localización por el usuario no le demande salir de la plataforma o demande navegar múltiples sitios web para encontrar las últimas versiones actualizadas de dichas políticas.
- Prever en sus "términos y servicios y reglas de comunidad, [que las empresas] no deben limitar o restringir la libertad de expresión de manera desproporcionada o innecesaria".

    - **Sugerimos:** Exigir a las empresas que expresamente se alineen a los Principios sobre Empresas y Derechos Humanos, a que adopten el Plan de Acción de Rabat que aporta elementos valiosos para la clasificación del discurso de odio, y que demuestren su apego a los estándares interamericanos elaborados por la CIDH y la RELE como fruto de esta consulta.
- No obstaculizar el acceso a contenidos sobre elecciones y la disponibilidad de diversos puntos de vista para los usuarios.

Tercero. Las obligaciones fruto del impacto de las prácticas de moderación:

- Adoptar medidas de transparencia sobre el uso e impacto que pueden ocasionar herramientas automatizadas de moderación de contenidos, incluyendo información sobre en qué medida afectan, entre otros, la clasificación o eliminación de los mismos.
    - **Sugerimos:** Obligar a las plataformas a publicar de manera transparente la tipología de discursos de odio que ha sido moderada por la red social, desagregando la información en razón a los discursos de odio en razón a su
- Deben establecer "sistemas eficaces de vigilancia, evaluaciones de impacto, sistemas de denuncias accesibles y eficaces a fin de identificar los daños reales o potenciales a los derechos humanos causados por sus servicios o actividades".
    - **Sugerimos**: Exigir a las plataformas de internet reestablecer o facilitar para investigadores, sociedad civil y usuarios el acceso a herramientas de investigación sobre los contenidos que circulan en línea, su moderación, alcance y audiencia, impacto, entre otros.
- Deben disponer de mecanismos adecuados para proporcionar remedios apropiados a los afectados.
    - **Sugerimos:** Integrar mecanismos de participación de múltiples partes para que éstas puedan llevar a cabo tareas de evaluación y seguimiento el impacto de la moderación de contenidos en la protección de la libertad de expresión.
- Deben ajustar sus actividades y sistemas para prevenir futuros abusos a los derechos de sus usuarios.

Citas

---

[i] Incluida la Declaración Conjunta sobre Libertad de Expresión y Elecciones en la Era Digital de 2020; la Declaración Conjunta del Vigésimo Aniversario para la Libertad de Expresión en la Próxima Década de 2019; la Declaración Conjunta sobre Libertad de Expresión y Noticias Falsas ("Fake News"), Desinformación y Propaganda de 2016; el Informe "Estándares para una Internet Libre, Abierta e Incluyente" de 2016; el Informe Libertad de Expresión e Internet de 2013; y la Declaración Conjunta de 2011 sobre Libertad de Expresión.

[ii] https://www.oas.org/es/cidh/expresion/showarticle.asp?artID=1056&lID=2 ver punto 4 c

[iii] Incluido el Convenio de Budapest, suscrito voluntariamente por varios países de América Latina; el Protocolo Adicional de la Convención de los Derechos del Niño sobre venta de niños, prostitución infantil y pornografía infantil; y el arto 19 de la Convención Americana sobre los Derechos Humanos. También la Relatoría Especial para la Libertad de Expresión señaló en su informe "Estándares para una internet libre, abierta e incluyente" que entre las modalidades de los discursos no protegidos se encuentran "la propaganda de guerra y apología al odio que incita a la violencia, la incitación directa y pública al genocidio, y la pornografía infantil".
https://www.oas.org/es/cidh/expresion/docs/publicaciones/INTERNET_2016_ESP.pdf Ver párrafo 78.

[iv] Si bien ni la CADH ni las interpretaciones de la CIDH o la RELE aportan definiciones claras sobre el discurso de odio, sí que aportan elementos para definir cuándo el discurso de odio merece ser sancionable por los Estados. En el informe de Violencia por Prejuicio contra personas LGBTI se usa la expresión "sancionable" más no ilegal, por lo

que está abierto a interpretación —-y es objeto de otros aspectos de esta consulta— si debería, cuando es sancionable, ser tratado por la vía penal, civil o administrativa.

[v] En el informe de Violencia por Prejuicio contra personas LGBTI, párrafos 229 y ss https://www.oas.org/es/cidh/informes/pdfs/ViolenciaPersonasLGBTI.pdf

[vi] En el informe de Violencia por Prejuicio la CIDH y la Rele sumaron como criterios protegidos la identidad de género, la diversidad corporal, y la orientación sexual. Ver párrafos 229 y siguientes https://www.oas.org/es/cidh/informes/pdfs/ViolenciaPersonasLGBTI.pdf

[vii] En el informe de Violencia por Prejuicio contra personas LGBTI, párrafos 232 y ss https://www.oas.org/es/cidh/informes/pdfs/ViolenciaPersonasLGBTI.pdf

[viii] En el informe de Violencia por Prejuicio contra personas LGBTI, párrafos 232 y ss https://www.oas.org/es/cidh/informes/pdfs/ViolenciaPersonasLGBTI.pdf

[ix] Ver https://www.derechosdigitales.org/wp-content/uploads/discurso-de-odio-latam.pdf pg. 20

[x] Ver resolución A/74/486 del 9 de octubre de 2019, párrafos 49 y ss https://documents.un.org/doc/undoc/gen/n19/308/16/pdf/n1930816.pdf?token=ARcDGlatJnsrdqnT2S&fe=true

[xi] Ver https://www.ohchr.org/es/freedom-of-expression#:~:text=El%20Plan%20de%20Acci%C3%B3n%20de%20Rabat%20sobre%20la,Ginebra%2C%20Viena%2C%20Nairobi%2C%20Bangkok%20y%20Santiago%20de%20Chile%29.

[xii] Ver https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/

[xiii] Ver: https://www.derechosdigitales.org/wp-content/uploads/discurso-de-odio-latam.pdf pg.18

[xiv] Ver https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/

[xv] Ver https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/

[xvi] Ver https://transparency.meta.com/es-la/policies/community-standards/hate-speech/

[xvii] Ver https://www.oversightboard.com/decision/fb-uk2rus24/ del OVB

[xviii] Ver https://rfob.medium.com/facebook-oversight-boards-transparency-report-hampered-by-meta-s-own-lack-of-transparency-4581b2e96c3d

[xix] Ver https://www.linternaverde.org/blog/adios-a-crowdtangle-el-nuevo-panorama-del-acceso-a-datos-en-facebook-e-instagram

[xx] Ver pg. 20https://www.derechosdigitales.org/wp-content/uploads/discurso-de-odio-latam.pdf

[xxi] https://santaclaraprinciples.org/

[xxii] https://www.alsur.lat/sites/default/files/2022-05/moderacion_contenidos_alsur.pdf

[xxiii] https://www.oas.org/es/cidh/expresion/showarticle.asp?artID=1056&lID=2 ver punto 4 c