unesco

Funded by the European Union

Moderación de contenidos y partes interesadas locales en Colombia

Marzo 2024

SOCIAL MEDIA 4PEACE

Publicado por ARTICLE 19, marzo de 2024

ARTICLE 19

72-82 Rosebery Ave

London EC1R 4RW

UK

T: +44 20 7324 2500
F: +44 20 7490 0566
E: info@article19.org
W: www.article19.org
Tw: @article19org
Fb: facebook.com/article19org

Esta publicación se realizó con el apoyo financiero de la **Unión Europea** y la **UNESCO**. Las denominaciones empleadas y la presentación del material en esta publicación no implican ninguna opinión por parte de la UNESCO o de la Unión Europea sobre la condición jurídica de ningún país, territorio, ciudad o zona, o de sus autoridades, ni respecto de la delimitación de sus fronteras o límites.

Los autores son responsables de la elección y presentación de los hechos contenidos en la publicación y de las opiniones expresadas en esta, que no son necesariamente las de la Unión Europea o la UNESCO y no comprometen a las Organizaciones.

Agradecimientos

Agradecemos especialmente a Carolina Botero Cabrera, Catalina Moreno Arocha y Darly Díaz, quienes realizaron la investigación y escribieron este reporte con el acompañamiento de ARTICLE 19.

Carolina Botero Cabrera es directora ejecutiva de Fundación Karisma y columnista de El Espectador y La Silla Vacía. Abogada, Máster en Derecho Internacional y Cooperación y Máster en Derecho del Comercio y Contratación. Ha trabajado durante más de una década en la promoción y defensa de los derechos humanos en Internet. Forma parte de la Junta Directiva de Creative Commons y del Consejo Asesor de la Sociedad Civil de la Información (CSISAC) de la OCDE.

Catalina Moreno Arocha es coordinadora de la Línea de Inclusión Social de la Fundación Karisma. Es abogada y tiene una maestría en Derecho Público. Se dedica a temas de derechos humanos y de justicia social. Trabajó durante más de una década en la Corte Constitucional colombiana y como abogada de incidencia política y jurídica en asuntos de género. En Karisma se dedica a promover que las tecnologías sirvan y protejan a grupos sociales que están expuestos a violencias y a discriminaciones.

Darly Díaz es una socióloga y periodista colombiana, Magister en Estudios de Desarrollo con énfasis en derechos humanos, perspectivas de la justicia social y género. Cuenta con más de cinco años de experiencia trabajando con diversas comunidades (afrocolombianos, periodistas y niños bajo medidas de protección) y en la defensa y promoción de los derechos humanos en ambientes digitales y físicos.

Muchas gracias a todos los que compartieron su experticia y sus opiniones para esta investigación. Agradecemos también a nuestros colegas de la UNESCO, quienes brindaron apoyo y retroalimentación en distintas fases de la redacción de este reporte.

Contenido

Resumen ejecutivo

Este informe examina las prácticas de moderación de contenidos de las empresas de redes sociales en Colombia y sus implicaciones en la libertad de expresión. Identifica desafíos significativos dentro del ecosistema informativo facilitado por las plataformas de redes sociales. Las prácticas de moderación de contenidos de las grandes plataformas en Colombia fueron analizadas detalladamente por las partes interesadas que participaron en esta investigación, resaltando aspectos claves.

Falta de transparencia

Hace falta más transparencia en las prácticas de moderación de contenidos y, aunque las plataformas ofrecen información sobre sus procesos, esta no es lo suficientemente clara en Colombia (así como sucede en otras partes del mundo). Las apelaciones por lo general no son respondidos y las partes interesadas consideran que no hay proporcionalidad ni consistencia en las decisiones de moderación de contenido. Mientras que las plataformas usualmente notifican a los usuarios cuando se modera el contenido, existe poca información sobre los mecanismos de apelación; y usualmente, estos se ven como ineficaces. Algunas plataformas ofrecen información sobre la moderación de contenidos en Colombia —a través de sus informes de transparencia—; sin embargo, esta información no siempre es útil para identificar patrones o entender la escala y el impacto de la moderación de contenidos en el país.

Esta falta de transparencia impide que la sociedad civil y los usuarios puedan entender, afecta negativamente los esfuerzos de incidencia y, en últimas, debilita la rendición de cuentas de las plataformas. La transparencia relacionada con las prácticas de moderación de contenidos es fundamental para que la sociedad civil comprenda cómo funciona la moderación de contenidos y cuál es su impacto en cada contexto particular. Por lo tanto, es importante interactuar con cualquier marco regulatorio que afecte la moderación de contenidos directa o indirectamente, como las obligaciones de transparencia, o que se vea influenciado por ella. Una creciente preocupación es la opacidad y percepción de censura que tienen las prácticas de curación de contenidos. Es

urgente entender de mejor manera las prácticas de moderación y curación de contenidos como la disminución de visibilidad (*downranking*) o la censura encubierta (*shadowbanning*), las cuales son menos visibles y estudiadas que la eliminación o suspensión de cuentas.

Falta de comprensión contextual

Aunque las redes sociales son un actor fundamental del ecosistema de la información, todavía existe una falta de comprensión contextual sobre el funcionamiento de la moderación y curación de contenidos en Colombia. Las plataformas se aproximan a la moderación de contenidos como un asunto global, a pesar de la importancia de entender los contextos locales. Si bien las plataformas son conscientes de esto y han implementado estrategias para aproximarse a los matices propios de algunos contextos, las personas entrevistadas consideran que estos esfuerzos son insuficientes.

Una de las consecuencias de esta falta de comprensión contextual es la exclusión de grupos de personas que son particularmente vulnerables en Colombia y que son sujetos de ataques en línea, como los defensores de derechos humanos y los periodistas, especialmente las mujeres periodistas. De esta manera, aumenta la probabilidad de que estas personas estén expuestas a violencia fuera de línea. También existe la dificultad de superar la moderación de contenidos para informar sobre asuntos de interés público, incluyendo las violaciones a derechos humanos. Esto puede suponer medidas como la prohibición a actores que participan activamente en procesos de paz con el gobierno o la eliminación de contenidos que denuncien abusos policiales durante protestas sociales, así como la censura de insultos que hacen parte de cánticos de protesta durante momentos políticos cruciales para el país.

Prácticas de moderación de contenido

Los desafíos derivados de las prácticas de moderación de contenidos también impactan en la relación entre los medios y las plataformas de redes sociales. Si bien las plataformas de redes sociales son un canal de distribución primario, depender de ellas

puede tener un impacto negativo en los medios debido a los desafíos de moderación de contenidos, lo que podría resultar en autocensura por parte de los medios al tomar decisiones editoriales.

El uso extendido de herramientas automatizadas de moderación de contenidos, aunque es comprensible su implementación, tiene el potencial de exacerbar todas las problemáticas de la moderación de contenidos descritos, y ha sido señalado como un problema por varios actores entrevistados.

Rol del Estado

La injerencia del Estado en la moderación de contenidos también ha suscitado preocupación. Las autoridades colombianas, que carecen de base legal para hacerlo, han realizado múltiples solicitudes a las plataformas para retirar contenidos teniendo en cuenta sus normas comunitarias. Existe una falta de transparencia en relación con estas solicitudes a las plataformas, lo que dificulta la comprensión de los motivos de las solicitudes, cómo son evaluadas por las plataformas a la luz de los estándares de derechos humanos en Colombia, y qué autoridades son las responsables de realizar las solicitudes.

Análisis y conclusiones

Teniendo en cuenta este contexto, el presente informe analiza la viabilidad de establecer una coalición local sobre moderación de contenidos y libertad de expresión en Colombia para promover la consolidación de canales de comunicación y cooperación con empresas de redes sociales y reguladores, y para tratar aquellas temáticas que amenazan la libertad de expresión en línea, la libertad de prensa y la cohesión social.

El informe concluye que, en el contexto colombiano, aprovechar las alianzas, redes y organizaciones preexistentes que trabajan en derechos digitales y libertad de expresión tendría más éxito que establecer una iniciativa nueva. El informe cierra con algunas recomendaciones que apoyan esta propuesta.

Introducción

Esta publicación hace parte del proyecto **Social Media 4 Peace** de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), financiado por la Unión Europea (UE).

Acerca del proyecto

Este informe hace parte del proyecto **Social Media 4 Peace** implementado por la UNESCO en Colombia, Bosnia y Herzegovina, Kenia e Indonesia con el apoyo de la Unión Europea. El objetivo general del proyecto es reforzar la resiliencia de las sociedades frente a los contenidos potencialmente perjudiciales difundidos en línea, en particular el 'discurso de odio' y la 'desinformación'[1], al tiempo que se protege la libertad de expresión y se contribuye a la promoción de narrativas de paz a través de las tecnologías digitales, especialmente las redes sociales. La contribución de ARTICLE 19 al proyecto se centra en las preocupaciones planteadas por las prácticas actuales de moderación de contenidos en las principales plataformas de redes sociales de los cuatro países objetivo.

Además de cuatro informes de país elaborados por consultores externos, ARTICLE 19 también publicó un informe resumen sobre Bosnia y Herzegovina, Indonesia y Kenia que comparaba los aprendizajes y recomendaciones.

ARTICLE 19 considera que las empresas de redes sociales tienen la libertad, en principio, de restringir el contenido que circula sobre la base de la libertad contractual, pero que deben hacerlo respetando derechos humanos, incluyendo los derechos a la libertad de expresión, la privacidad y el debido proceso. Si bien las plataformas de redes sociales han ofrecido oportunidades para la expresión, diversos asuntos problemáticos han salido a la luz. La aplicación de normas comunitarias ha llevado al silenciamiento de las voces de las minorías. Los esfuerzos de las empresas de tecnología para hacer frente a los contenidos problemáticos distan mucho de estar distribuidos uniformemente. Por ejemplo, en 2021 se reportó que el 87% del gasto de Facebook en desinformación se asignaba a contenidos en inglés, a pesar de que solo el 9% de sus usuarios eran anglófonos. La información

filtrada también reveló que la mayoría de los recursos de moderación de contenidos se destinaban a un número limitado de países. Al mismo tiempo, la transparencia y la resolución de controversias sobre remover contenidos han sido insuficientes para favorecer el escrutinio de las acciones de las plataformas de redes sociales y para ofrecer una reparación significativa a sus usuarios. Finalmente, es preocupante que un reducido número de plataformas dominantes tenga tanto poder sobre lo que las personas pueden ver sin una rendición de cuentas pública más directa.

Este informe se centra en la situación de los actores locales en Colombia. La investigación realizada revela que, si bien la circulación de contenido potencialmente 'perjudicial' en redes sociales o la moderación del mismo impacta a ciertos actores, estos por lo general son incapaces de tomar medidas efectivas para mejorar su situación en este asunto. En algunos casos, se sienten frustrados por las inconsistencias en la manera en la que las plataformas aplican sus propias normas comunitarias. En otros casos, sienten que las plataformas ignoran sus solicitudes o malinterpretan las circunstancias y contextos específicos del país o la región. Algunos actores no comprenden las normas o la moderación de contenidos en general.

Esta investigación examina las perspectivas de las partes interesadas sobre este asunto. Asimismo, se pregunta por el papel que podría desempeñar una coalición sobre moderación de contenidos y libertad de expresión para mejorar las condiciones y la materialización de los derechos en la era digital. Busca también orientar sobre las mejores formas y estrategias para establecer conexiones con el fin de cerrar la brecha entre las realidades de los actores locales, el sector público y las empresas privadas que operan a escala global en moderación de contenido.

La idea de coaliciones nacionales se basa en la premisa de que para las plataformas de redes sociales es esencial adquirir una comprensión del contexto local en el que operan y colaborar con los actores locales. Recopilar conocimiento local y entender el contexto (lingüístico, histórico, político y social) les permitiría a las plataformas de redes sociales mejorar sus prácticas de moderación de contenidos y hacerlas contextualmente

apropiadas. Una coalición local sobre libertad de expresión o alguna estructura similar podría interactuar en un diálogo productivo con las plataformas de redes sociales y contribuir a que se atiendan los desafíos en la moderación de contenidos y se fortalezca la protección de los derechos fundamentales en línea. Además, podría participar en el desarrollo de capacidades proporcionando formación y apoyo sobre moderación de contenidos y libertad de expresión a otros agentes locales de la sociedad civil afectados por la moderación de contenidos.

Durante esta investigación, se propuso a distintos actores interesados la idea de crear una coalición local sobre moderación de contenidos y libertad de expresión en Colombia. Sus opiniones permitieron realizar recomendaciones sobre cómo la propuesta de coalición podría abordar los problemas de moderación de contenidos en Colombia.

Con eso en mente, y centrándose en las voces locales en Colombia, este informe examina los asuntos de moderación de contenidos desde una perspectiva local, incluyendo estudios de caso, y la posición, conocimiento y necesidades de distintos actores estatales y no estatales. Se resalta la diversidad y complejidad de la sociedad e historia colombianas como trasfondo para entender el informe. También se presenta cómo los profundos conflictos de la sociedad se han explotado, en ocasiones, para obtener beneficios políticos y económicos.

Este informe inicia describiendo el panorama de las redes sociales y explorando las dinámicas y asuntos relacionados con el uso de redes sociales y las prácticas de moderación de contenidos en el país. Luego discute cómo crear una coalición sobre moderación de contenidos y libertad de expresión y examina las necesidades, brechas y fortalezas de una posible coalición. A continuación, analiza los grupos de partes interesadas que se ocupan de las prácticas de moderación de contenidos o se ven afectados por ellas. El informe concluye con recomendaciones sobre la viabilidad de la formación de una alianza de la sociedad civil sobre moderación de contenidos y libertad de expresión en Colombia para tender un puente de diálogo entre las redes sociales y la sociedad civil local.

En este informe, nos basamos en las siguientes definiciones:

- **Moderación de contenidos** incluye las distintas medidas y herramientas que una plataforma de redes sociales usa para hacer frente a los contenidos ilícitos y para aplicar sus normas comunitarias a los contenidos generados por los usuarios. Esto generalmente implica la denuncia por parte de los usuarios, 'trusted flaggers', o 'filtros'; la remoción, el etiquetado, la disminución de visibilidad (*downranking*) o desmonetización del contenido, o la desactivación de ciertas funciones.
- **Curación de contenidos** se refiere a cómo las plataformas de redes sociales usan sistemas automatizados para clasificar, promocionar o disminuir la visibilidad (*downranking*) del contenido en el *feed* de noticias, normalmente basándose en los perfiles de sus usuarios. El contenido también puede ser promovido en las plataformas a cambio de un pago. De igual modo, las plataformas pueden curar contenidos a través de anuncios intersticiales para advertir a los usuarios sobre cierto contenido sensible o aplicando etiquetas para destacar, por ejemplo, si el contenido proviene de una fuente confiable.

Ambos procesos pueden sobreponerse. Por ejemplo, disminuir la visibilidad (*downranking*) de cierto contenido puede ser una medida de la moderación de contenido, pero también una parte inevitable del proceso de curación de contenidos.

Metodología

En este informe se usaron varias metodologías de investigación. Primero, la investigación se basó en una revisión exhaustiva de la literatura académica y no académica para ofrecer una visión comprensiva de los asuntos a tratar. Posteriormente, la recolección de datos cualitativos permitió recoger las distintas perspectivas de diversos actores. Se organizaron entrevistas con el fin de entender las experiencias locales y los desafíos a los que se enfrentan al tratar con las plataformas en asuntos de moderación de contenido.

Se desarrollaron cuatro cuestionarios distintos para guiar las entrevistas con plataformas de redes sociales, organizaciones de la sociedad civil, medios de comunicación y la

academia. Algunas preguntas se aplicaron de forma general, mientras que otras se personalizaron en función de la proximidad y el papel del entrevistado en los procesos de moderación de contenido. Algunos cuestionarios se referían a la opinión de los entrevistados sobre contenido potencialmente 'perjudicial' y a cómo podrían imaginar una coalición que trabajara conjuntamente en estos temas.

En total, se realizaron 23 entrevistas y se recibieron 5 contribuciones escritas (ver Anexo B). Las investigadoras se contactaron con un mayor número de actores, pero no recibieron respuesta de todos los contactados. La mayoría de entrevistas se realizaron a través de Zoom y algunas contribuciones se recolectaron a través de una encuesta en línea que se compartió con los actores que no tuvieron tiempo de participar en una entrevista. La encuesta también se publicó en las redes sociales de la Fundación Karisma para incluir otras voces de partes interesadas.

Cada entrevista y encuesta explicaba el propósito del proyecto, la relevancia de los participantes y cómo sus respuestas serían tratadas. Karisma distribuyó formularios de consentimiento por escrito que fueron firmados por los participantes. En algunos casos, las investigadoras solicitaron y recibieron el consentimiento verbal de los entrevistados al inicio de la entrevista para utilizar la información recolectada en este informe.

Aunque se contactó a algunas instituciones gubernamentales relacionadas con los asuntos de moderación de contenido, no fue posible obtener una entrevista. El Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) solo facilitó comentarios por escrito. Las investigadoras también intentaron, sin éxito, que la Fiscalía General de la Nación y la Procuraduría General de la Nación participaran en la investigación. A petición de la UNESCO, el Ministerio de Asuntos Exteriores se puso en contacto con el Ministerio de Defensa Nacional y la Superintendencia de Industria y Comercio.

Las organizaciones de la sociedad civil entrevistadas trabajan en una amplia gama de asuntos: protección y defensa de los derechos de la infancia, organizaciones de mujeres, organizaciones e individuos trans, colectivos en favor de los derechos de las personas

afrocolombianas, red nacional indígena, libertad de expresión, libertad de prensa, lucha contra la desinformación en línea, construcción de paz y sensibilización sobre el uso de sustancias psicoactivas. También se entrevistó a periodistas, académicos y centros de estudio.

Las entrevistas con Meta, X (antes Twitter) y Google contaron con la participación de varios representantes de las plataformas, por lo que no se trató de una reunión individual, sino de sendas conversaciones con el equipo de diversas dependencias al interior de las compañías.

Colombia en breve

Para entender la dinámica de moderación de contenidos y libertad de expresión en Colombia es importante primero entender el complejo sistema político y de gobierno del país. La estructura de gobierno actual se estableció en la Constitución Política de 1991. Según la Comisión de la Verdad[2], la proclamación de la nueva Constitución fue un punto de inflexión en la historia del país.[3]

Antes de la entrada en vigor de la nueva Constitución, y bajo un esquema de distribución bipartidista del poder, se excluyeron a distintos movimientos políticos entre 1958 y 1977. Como resultado del descontento social, surgieron guerrillas de ideología de izquierda y resistencia armada. El descontento se agudizó entre 1978 y 1991, el levantamiento armado se consolidó y hubo una fuerte respuesta represiva por parte del Estado colombiano. La guerra contra las drogas también comenzó. Fue una época de permanente estado de sitio, con un aumento de las violaciones a los derechos humanos. Este periodo terminó con la convocatoria a una Asamblea Nacional Constituyente en 1991, una iniciativa promovida por los movimientos sociales, especialmente los estudiantes, dentro de los diálogos con la guerrilla del M-19.

La nueva Constitución acabó el anterior sistema político: propuso un modelo democrático más pluralista e inclusivo. Para la Comisión de la Verdad, sin embargo, el impacto de la

Constitución ha sido desigual en distintas regiones y grupos poblacionales. La Comisión también estableció que hubo una reacción violenta a la apertura democrática que trajo la nueva Constitución porque dos de los grupos armados más importantes del país (el Ejército de Liberación Nacional —ELN— y las Autodefensas Gaitanistas de Colombia) no fueron incluídos en este nuevo pacto y las negociaciones con los líderes de los cárteles de droga fracasaron incluso antes de haber terminado. Al mismo tiempo, el movimiento por la paz ganó fuerza y logró desescalar el conflicto armado, aunque en 2023 el conflicto todavía no había terminado.

Según la Constitución de 1991, Colombia es una república unitaria, con descentralización administrativa y una distribución del poder entre el gobierno nacional y los gobiernos locales. Existen tres ramas del poder público: legislativo, ejecutivo y judicial (con un mecanismo de 'frenos y contrapesos') y ciertos organismos de control autónomos con funciones especializadas. Otro mecanismo judicial, la *acción de tutela* ante la Corte Constitucional, es de gran importancia para los derechos fundamentales.

Cualquier ciudadano puede interponer una *tutela* con unos requisitos mínimos en materia probatoria y sin ser abogado o tener conocimientos legales. Esta debe resolverse dentro de los diez días siguientes a su recepción y sirve para la protección de cualquier derecho fundamental, incluyendo el derecho a la libertad de expresión, a la libertad de prensa y a la participación política. Su uso es cada vez más popular: en 1992, cuando se introdujo este mecanismo, se presentaron 10.000 tutelas; en 2022, se presentaron 633.463 tutelas. La tutela transformó la forma en la que se entendió la ley y facilitó que los ciudadanos se apersonaran de sus derechos. La mayoría de los casos de la Corte Constitucional sobre libertad de expresión han sido producto de fallos de tutela.

La Constitución protege la libertad de expresión y la libertad de prensa. El artículo 20 garantiza la libertad de expresar y difundir ideas y opiniones, así como el de recibir información veraz e imparcial. El artículo 20 también garantiza el derecho a establecer medios de comunicación y asegura el derecho de rectificar información publicada en condiciones de equidad. También indica que 'no habrá censura'.

A pesar del nuevo andamiaje constitucional, dos ideas que habían servido para estigmatizar a los movimientos sociales y a miembros de la oposición política —la 'doctrina de la seguridad nacional' y la idea del 'enemigo interno'— se fortalecieron en los manuales y en la regulación del Ejército relativos al combate contrainsurgente. Estas ideas legitimaron las acciones violentas realizadas por el Estado contra miembros de la oposición, líderes estudiantiles, rurales y sociales.[4]

La estigmatización de defensores de derechos humanos y oponentes políticos continuó bajo la vigencia de la nueva Constitución. A inicios de los 2000, los ataques verbales de políticos de alto rango contra defensores de derechos humanos y periodistas aumentaron. Concretamente, se intensificaron durante la presidencia de Álvaro Uribe Vélez. Por ejemplo, en 2003 y 2004 organizaciones de la sociedad civil presentaron acciones de tutela contra el entonces presidente Álvaro Uribe Vélez por sus declaraciones públicas demonizando a los defensores de derechos humanos.[5]

Esto también ha sido resaltado por la Corte Constitucional colombiana, que ha declarado que aquellos en posiciones de poder tienen un grado de responsabilidad frente al público y deben garantizar que sus declaraciones y discursos públicos se ajusten a los límites del derecho a la libertad de expresión.[6] Esta Corte recordó que los criterios de veracidad e imparcialidad de la información, justificación fáctica y razonabilidad de las opiniones y respeto por los derechos fundamentales de los ciudadanos son aspectos esenciales al analizar los discursos de los servidores públicos.[7]

Las organizaciones de la sociedad civil continuaron litigando contra las declaraciones realizadas por servidores públicos en altos cargos cuando estas promovieran ideas negativas sobre grupos vulnerables como las personas migrantes[8] o las mujeres periodistas[9], o cuando se refirieran al ejercicio de derechos fundamentales, como los derechos sexuales y reproductivos. La estigmatización de los defensores de derechos humanos todavía es visible en Colombia en el mundo digital y pone en riesgo el trabajo y la seguridad de activistas y organizaciones de la sociedad civil.

Vale la pena mencionar que entre 2019 y 2021 Colombia experimentó los momentos más importantes de protesta en su historia reciente. Las protestas se originaron en la propuesta presentada por el gobierno ante el Congreso de la República de una reforma tributaria y una reforma educativa. Sin embargo, las protestas aumentaron en 2021 debido al descontento social relacionado con la reaparición de la pandemia por Covid-19. El gobierno tomó un camino basado en el orden público para hacer frente a las protestas y las demandas sociales no se tramitaron a través de mecanismos o canales institucionales.

Luego de una visita de trabajo para investigar las violaciones a los derechos humanos cometidas durante las protestas sociales en 2021, la Comisión Interamericana de Derechos Humanos (CIDH) indicó que:

> *la existencia de un clima de polarización que se relaciona de forma directa, tanto con la discriminación estructural étnico racial y de género, como con factores de carácter político. Este fenómeno está presente en diferentes sectores sociales y se manifiesta en discursos estigmatizantes que a su vez propician un acelerado deterioro del debate público. La Comisión Interamericana encuentra especialmente preocupante estos discursos cuando provienen de autoridades públicas.* (párr. 5)

Para la CIDH, Internet ha permitido a las personas manifestantes en Colombia comunicar incidentes y hacer denuncias abiertas sobre el uso excesivo de la fuerza, además de solicitar la protección de sus derechos, facilitando y enriqueciendo la deliberación pública y denunciando las violaciones a los derechos humanos durante las manifestaciones. Esto ha evidenciado la necesidad de garantizar el libre acceso a Internet. La CIDH recibió denuncias sobre presuntas medidas estatales que podrían cercenar las libertades en línea, como el ciber-patrullaje y las actividades de perfilamiento, la clasificación de contenido en línea como verdadero o falso por las autoridades de policía, cortes de internet y bloqueo de direcciones IP. La CIDH estableció que 'según la información entregada por distintos actores, dichas acciones estarían siendo emprendidas por

criterios subjetivos en vez de parámetros objetivos, legítimos y transparentes, conforme a estándares internacionales de derechos humanos' (párr. 174).

La CIDH también indicó que la mayoría de las partes interesadas entrevistadas durante su visita afirmaron que, aunque Internet es una plataforma importante para la deliberación pública, también 'manifestaron temores de que algunos discursos incentiven la violencia o sean la base para la toma de decisiones sobre Internet que resten voz a quienes quieren expresarse sobre asuntos de interés público' (párr. 175).

Entre 2002 y 2016, el Estado respondió a los grupos insurgentes a través de la acción militar. Dicho conflicto causó serias violaciones a los derechos humanos en tanto los grupos combatientes afectaron, directa e indirectamente, a la población civil. La justicia transicional empezó en 2005 con un acuerdo de paz entre el gobierno y los grupos paramilitares de ideología de derecha. Luego de las negociaciones entre el gobierno y las Fuerzas Armadas Revolucionarias de Colombia – Ejército del Pueblo (FARC-EP), se firmó el Acuerdo de Paz en 2016.

A pesar del Acuerdo, la paz está lejos de ser una realidad en Colombia. Hoy el país se enfrenta a 'una serie de confrontaciones regionales fragmentadas que, aunque no están del todo desconectadas entre sí, a diferencia de las décadas anteriores, no tienen como columna vertebral la disputa por el poder político ni el control del Estado'[10]. Mientras se implementan algunas reformas estructurales propuestas en el Acuerdo de Paz, el país se enfrenta también al problema del tráfico de drogas y sus rentas ilegales que nutren la violencia. El asesinato de líderes sociales y defensores de derechos humanos, así como de ex-combatientes de las FARC-EP, ha aumentado. Según un estudio realizado en 2021 por Movilizatorio, un laboratorio de transformación social, los asuntos relacionados con el Acuerdo de Paz y la Jurisdicción Especial para la Paz[11] se encuentran entre los temas más polarizantes en Colombia, al menos en X.

La polarización no es nueva en Colombia; se puede rastrear a través de distintos momentos de la violenta historia política del país. El estudio de Movilizatorio encontró que las redes sociales contribuyen a la intensificación de la polarización porque permiten la

difusión rápida de cierto contenido y la autoafirmación de un discurso como resultado de los sistemas de recomendación que usan las plataformas. Sin embargo, el análisis de Movilizatorio también concluyó que 'a pesar de la percepción generalizada de polarización, y de la polarización que se genera alrededor de ciertas agendas temáticas, existe en Colombia una unidad de valores morales sobre la cual es posible construir acuerdos para el país'.

Por otro lado, según el Índice de *Freedom in the World Index*, que evalúa la condición de los derechos políticos y las libertades civiles alrededor del mundo, Colombia se consideró como un país libre en 2023. Este índice argumenta que, a pesar de una campaña polarizada, las elecciones de 2022 fueron libres y justas (el puntaje del reporte es 4/4). El puntaje en el indicador de estado de derecho fue 3/4 porque 'el sistema de justicia sigue comprometido por la corrupción y la extorsión', incluyendo el hecho de que 'la Corte Constitucional ha sido la encargada de mediar disputas políticas polarizantes múltiples veces, en especial relacionadas con la Jurisdicción Especial para la Paz, un sistema de justicia paralelo que se encuentra en el centro del sistema de justicia transicional creado por el Acuerdo de Paz de 2016'.

La instrumentalización de la polarización política en redes sociales incluye campañas de 'desinformación', como aquellas relacionadas con el plebiscito para la paz.[12] Durante las protestas de 2021, la respuesta del gobierno incluyó una campaña dirigida por el Ministerio de Defensa para identificar 'noticias falsas' durante las protestas en Colombia. Esto llevó a la limitación indebida de la libertad de expresión y ha contribuido a la difusión de 'desinformación'.[13] Por ejemplo, los medios de comunicación masiva han sido cooptados por grandes conglomerados políticos y económicos y se enfrentan a un entorno hostil como resultado de las críticas que reciben los periodistas de políticos y otras figuras públicas. En la zona rural, la población recibe información sesgada debido a la falta de acceso a medios de comunicación[14] o tiene experiencias de vida distintas debido a la presencia de actores armados que dominan su territorio.

El estado de la moderación de contenidos en Colombia

Panorama de las redes sociales en Colombia

En 2021, la población de Colombia era de 51,26 millones de personas: el 50,9% mujeres y el 49,1% hombres, con una edad promedio de 31,2 años. El 18% de la población vivía en áreas rurales y el 11,1% pertenecía a una minoría étnica.

La penetración de Internet ha incrementado de manera constante desde la pandemia de Covid-19, Según la Comisión de Regulación de Comunicaciones, la tasa de penetración de internet móvil era de 75,8 por cada 100 habitantes en septiembre de 2022 —5,8 puntos porcentuales más que en septiembre de 2021— mientras que la tasa de penetración de internet fijo era de 49,3 por cada 100 habitantes.

Según el *Boletín del primer trimestre de 2022* del MinTIC, a la fecha 8,52 millones de personas tenían acceso a internet fijo y 39,2 millones tenían acceso a conexiones móviles (13,8% a través de conexión 3G y 83,9% a través de conexión 4G).

Según el reporte anual Digital 2022: Colombia, en enero de 2022 había 35,5 millones de usuarios conectados a Internet en Colombia, con una tasa de penetración de 69,1%. El análisis realizado por Kepios muestra que el número de usuarios de Internet en el país se incrementó en 770.000 entre 2021 y 2022 (lo que equivale a un aumento de 2,2%). Esto quiere decir que el 30,9% permanece fuera de línea. Según datos de GSMA Intelligence, a inicios de 2022 existían 65,7 millones de conexiones móviles en Colombia.

El informe Digital 2022: Colombia muestra que a inicios de 2022 había 36,2 millones de usuarios de redes sociales mayores de 18 años en Colombia, lo que equivale al 93,9% de la población mayor de 18 años. De hecho, a enero de 2023 el 97,7% de las personas que hacen uso de Internet también hacen uso de al menos una plataforma de redes sociales, sin importar su edad. Un análisis de Kepios y We are Social mostró que el número de usuarios de redes sociales incrementó en 2,8 millones (+7,2%) entre 2021 y 2022. En enero de 2022, había 41,8 millones de cuentas de redes sociales, lo que representaría al

81% de la población. Cabe señalar que las cifras de usuarios se basan en cuentas de usuarios activos y no necesariamente se refiere a individuos particulares.

Cada día, los colombianos dedican 10 horas y 3 minutos a navegar en Internet en cualquier dispositivo, de las cuales 3 horas y 46 minutos las dedican a navegar en redes sociales. Esto pone al país de cuarto a nivel mundial. El mismo informe indica que las plataformas más usadas son WhatsApp (94%), Facebook (91,7%), Instagram (84,4%), Facebook Messenger (73,8%), TikTok (69,5%), y X (50,8%). El grupo Meta es dueño de tres de las cinco plataformas con más usuarios activos en el país (Facebook, Facebook Messenger, e Instagram). A pesar de que otras plataformas han aumentado su presencia, como Pinterest y TikTok, Meta sigue a la delantera en cuanto a tráfico en línea.

Según el informe, el mayor incremento en el uso de plataformas se dio entre personas de 25 a 34 años, siendo las mujeres el 14,8% y los hombres el 14,9%, quienes en su mayoría usan Facebook e Instagram para 'mantenerse en contacto con amigos y familia' (párr. 53). Los colombianos de este grupo etario usan redes sociales no solo como fuente de entretenimiento, sino por motivos educativos y de comunicación.

En Sudáfrica, Filipinas y Brasil las personas dedican más tiempo a usar Internet, estando Colombia en el cuarto puesto a nivel mundial (párr. 27). El motivo más usual para usar redes sociales es la búsqueda de información (párr. 29), como noticias o contenido político, y X destaca entre las otras plataformas en este tipo de uso. Lo anterior podría deberse a distintos factores como la inexistencia de infraestructura de servicios públicos en zonas rurales o a las políticas y regulaciones del *zero-rating* que permiten que los usuarios accedan sin costo a cierto contenido o aplicaciones —por lo general Facebook y WhatsApp— sin que este acceso consuma datos del plan de internet móvil contratado. También puede ser reflejo de un cambio generacional en el que las personas prefieren usar plataformas digitales.

Un informe publicado en 2022 por el Reuters Institute encontró una muy alta tasa de uso de telefonía móvil en una muestra urbana. Según el informe, 'las muestras tomadas en entornos digitales tienden a subrepresentar los hábitos de consumo de noticias de las

personas que son mayores y menos adinerados, lo que quiere decir que el uso de medios digitales se suele sobrerrepresentar y el uso de medios tradicionales fuera de línea subrepresentar. En ese sentido, es mejor pensar en los resultados como representativos de la población que se encuentra en línea'. Además, a medida que la población es más urbana, su fuente de noticias tiende a ser más en línea (86% incluyendo redes sociales) y menos la televisión (55%) o los medios impresos (28%). TikTok creció especialmente entre los jóvenes. Según el informe, estos números se deben a la aceleración en la digitalización ocasionada por la pandemia.

Sobre la búsqueda de noticias, el informe del Reuters Institute destaca que el 60% de la población accede a noticias en línea a través de redes sociales, 35% escribe una palabra clave o el nombre de una página de internet en un motor de búsqueda y el 27% accede a través de los motores de búsqueda usando palabras que se refieren a hechos noticiosos en particular. Solo el 27% de las personas encuestadas reportó navegar en la página de web o aplicación móvil de un medio de comunicación. Sin embargo, el 61% de los encuestados indicó que les preocupaba distinguir entre información real e información falsa en línea.

Según el informe, el consumo de noticias en general disminuyó ligeramente después de la pandemia. En medio de ella, se buscaba información constantemente en internet, pero luego de los picos de contagios y a medida que los programas de vacunación avanzaron, las personas se desinteresaron o se cansaron de la información relacionada con el Covid-19. Los medios de comunicación se enfrentaron al reto de ofrecer información de interés en formatos distintos y al consumo de redes sociales u otras plataformas de entretenimiento. El informe encontró también que los debates electorales habían afectado la encuesta. En consecuencia, los temores de desinformación giraban en torno a cuestiones políticas y se descubrió que '[los memes se han convertido en una forma popular de expresión política en redes sociales]'.

Aparte del informe del Reuters Institute, no existe ningún reporte que evalúe la confianza que los ciudadanos colombianos depositen en las noticias que encuentran en redes

sociales o que muestre el impacto que estas tienen en asuntos políticos o sociales. Sin embargo, existen algunas encuestas que muestran la confianza que poseen algunas poblaciones, como los jóvenes, en los medios de comunicación. Por ejemplo, el Sexto estudio de percepción de jóvenes realizado en 2023 con la participación de personas entre 18 y 32 años encontró que el 36% de los encuestados confía en las redes sociales, el 28% en los medios de comunicación y el 15% en los influenciadores digitales. Por el contrario, el 62% no confía en las redes sociales, el 71% no confía en los medios de comunicación y el 82% no confía en los influenciadores digitales.

Panorama general: Impacto de la moderación y la curación de contenidos en los conflictos y los derechos humanos

En Colombia, existen factores que deterioran el debate público y pueden promover la aparición de contenido en línea que impacta negativamente los derechos humanos, como la discriminación estructural, la creciente falta de confianza en los medios de comunicación y otros asuntos políticos. Esto sucede a pesar de la orden constitucional de ampliar las garantías constitucionales y los mecanismos para hacer realidad los derechos fundamentales, así como de múltiples esfuerzos por lograr la paz.

En la sociedad colombiana persisten dinámicas de discriminación estructural por motivos de género, orientación sexual o identidad de género, raza, discapacidad, origen nacional, entre otros. La polarización también persiste en torno a los esfuerzos por alcanzar la paz. La discriminación y polarización son comunes en la circulación de contenido en redes sociales en Colombia, algunas veces incluso vulnerando los derechos fundamentales de las personas e impactando negativamente en la paz y la estabilidad. Para entender de mejor manera el impacto de la moderación y curación de contenidos en la paz y estabilidad en Colombia es necesario explorar algunos casos de ‘desinformación’ y de violencia basada en género en línea que circularon en redes sociales en el país.

Desinformación

La 'desinformación' es un problema real y las tácticas usadas para difundir dicho contenido se han sofisticado. Algunas organizaciones electorales, como la Misión de Observación Electoral (MOE), han declarado que la información electoral y la publicidad política en Colombia han cambiado notoriamente desde el surgimiento de las redes sociales y de la posibilidad de 'segmentar, perfilar y medir la reacción de las audiencias ante determinadas acciones de comunicación'. La MOE ha evidenciado que el contenido se hace viral sin que los usuarios necesariamente evalúen su veracidad, especialmente cuando se alinea con sus opiniones y preferencias. Una pieza periodística realizada por France24 demostró que en la campaña para las elecciones presidenciales de 2022, las estrategias de 'desinformación' fueron más sofisticadas. *Fact-checkers* en el país, como ColombiaCheck, advirtieron que los votantes habían sido blanco de montajes cada vez más 'refinados', videos manipulados 'milimétricamente' e incluso usuarios que se hacían pasar por ellos. En estos escenarios, "la información falsa y engañosa tiene el potencial de 'radicalizar posiciones' sobre los candidatos y 'desorientar' a los votantes aprovechando 'la creciente desconfianza' en las autoridades electorales.[15]

El Global Disinformation Index (GDI) indicó que la difusión de 'desinformación' tiene consecuencias disruptivas y perturbadoras en Colombia. La evaluación de riesgo que realizó GDI en el mercado de noticias en Colombia encontró que la mayoría de dominios web (44%) se encuentran en riesgo medio, el 41% de los dominios en riesgo bajo y el 12% tiene un alto riesgo de desinformación. En general, la clasificación que obtienen los dominios web disminuye por deficiencias operativas, especialmente en lo que respecta a la falta de transparencia sobre la propiedad del sitio y su estructura de financiación. Asimismo, otras políticas operativas y editoriales, como las directrices de atribución de fuentes y las prácticas de *fact-checking* también impactan en la clasificación.

Violencia basada en género en línea

Sobre la violencia basada en género en línea, se registraron algunos actos de violencia digital contra candidatas al Congreso y a la Presidencia durante las elecciones de 2022.

Por ejemplo, Francia Márquez, candidata a la Vicepresidencia, fue víctima de múltiples comentarios sexistas, racistas y clasistas en redes sociales. Según la MOE, que denuncia públicamente todo comentario racista, sexista o clasista que afecte los derechos fundamentales de las personas, este tipo de discurso contra mujeres candidatas se repite sistemáticamente. Además, su alcance se ha ampliado en distintos medios, así como en redes sociales, con el fin de afectar negativamente el ejercicio de los derechos políticos de mujeres y personas racializadas.

En un informe de 2022, la MOE advirtió sobre el incremento de los actos de violencia contra lideresas sociales. El informe identificó vulneraciones específicas contra mujeres lideresas por motivo de su género. Como se señaló, contrario a lo que les ocurre a hombres líderes sociales, a quienes se les amenaza directa y exclusivamente a ellos, en el caso de las mujeres lideresas las amenazas por lo general incluyen referencias a su condición de mujeres y amenazas a sus personas cercanas.

La violencia basada en género en línea se inserta en una estructura patriarcal de la sociedad más amplia. Se encuentra documentado que en Colombia las mujeres sufren altos niveles de violencia. El hecho de que esto ocurra en Internet no disminuye la seriedad de esta violencia. El impacto de la violencia puede acentuarse según la notoriedad de la víctima, por el trabajo que hace o por su pertenencia a ciertos grupos, como los periodistas o las mujeres candidatas.

Este tipo de violencia busca reducir la participación y visibilidad de las mujeres y de las agendas y temas que promueven en Internet, así como impedir su participación en el debate público. Karisma ha encontrado que la violencia sexista contra periodistas está muy extendida, se dirige contra los cuerpos, la apariencia, el tono de voz, las habilidades profesionales y las capacidades de mujeres periodistas y comunicadoras. Los múltiples y sistémicos patrones del patriarcado producen un continuo de dominación masculina violenta que se ha normalizado. La violencia psicológica, sexual y el acoso sexual ocurre donde las periodistas trabajan y los agresores actúan con impunidad en los espacios físicos y digitales.

En particular, existen llamados para erradicar el abuso y la estigmatización dirigidos contra quienes intentan esclarecer la verdad histórica sobre el conflicto armado y las expresiones de violencia y graves violaciones a los derechos humanos. Indepaz, una organización de la sociedad civil dedicada a asuntos de paz, sostuvo en una entrevista realizada para esta investigación que existe una relación entre la estigmatización y el 'discurso de odio' por parte de la clase política, por un lado, y la persistencia y reconfiguración de la violencia armada en las zonas más alejadas del país, por el otro.

Regulación de 'discurso de odio', 'desinformación' y violencia basada en género en línea

El sistema jurídico colombiano trata los asuntos relacionados con el 'discurso de odio', la 'desinformación' y la violencia basada en género en línea solo de manera parcial. Según el artículo 13 de la Convención Americana sobre Derechos Humanos[16] —que hace parte del bloque de constitucionalidad y sirve como guía interpretativa de los derechos y deberes constitucionales— el 'discurso de odio' debe considerarse como una ofensa punible. En ese contexto, la Corte Constitucional colombiana ha enfatizado que 'para que el contenido de un mensaje pueda considerarse un discurso que incita al odio no es suficiente con que el mensaje emita un reproche sobre una conducta, o que resulte ofensivo para el sujeto reprochado. Es necesario también que el contenido del mensaje incite al odio o a la violencia, o a cometer algún hecho ilícito en contra del sujeto pasivo del mensaje'.

Ni el sistema jurídico colombiano ni el derecho internacional ofrecen una definición legal de 'desinformación'. A pesar de que el concepto fue tratado en la Sentencia T-627 de 2012, la Corte Constitucional no lo definió y se limitó a indicar que los servidores públicos deben guiarse por los deberes de veracidad e imparcialidad de la información al hacer pronunciamientos públicos, así como seguir los principios de justificación fáctica y razonabilidad de sus opiniones y de respeto por los derechos fundamentales de los ciudadanos.

No existe legislación específica que regule la violencia basada en género en línea en Colombia. Sin embargo, una ley sobre prevención y atención de la violencia contra la

mujer y ciertas conductas típicas recogidas en el Código Penal pueden servirle a las autoridades para combatir dicho fenómeno. El uso de la tutela como mecanismo para proteger los derechos a la imagen y a la privacidad, también puede servir ante la inexistencia de normas más específicas sobre violencia basada en género en línea, como lo propuso recientemente la Corte Constitucional. Estos mecanismos permiten que las mujeres soliciten medidas de protección que oscilan entre el acompañamiento psicológico y la eliminación del contenido, lo que puede estar restringido bajo los estándares internacionales de libertad de expresión. Vale la pena resaltar que la Corte Constitucional ha reconocido que se trata de un asunto de gran importancia y ha exhortado en dos ocasiones al Congreso para que lo regule (en las sentencias T-280 de 2022 y T-087 de 2023).

Si bien las leyes formales pueden influenciar la moderación de contenido, la 'regulación' más importante para esta moderación, son las normas comunitarias de las empresas de redes sociales. En términos sencillos, estas normas comunitarias suelen sentar los tipos de contenido que se permiten o prohíben. Los límites de cada plataforma pueden ser más estrictos que los estándares de derechos humanos a nivel nacional, regional o internacional.[17] También puede suceder que las normas comunitarias y cómo se hacen cumplir no respondan adecuadamente a los discursos que circulan en redes sociales. En estas, la eliminación de contenido legítimo y la no eliminación de expresiones que afectan a personas o poblaciones en particular pueden tener consecuencias negativas para el debate público en línea.

Si bien durante las entrevistas realizadas para esta investigación las plataformas indicaron que la libertad de expresión es una prioridad en su trabajo, la siguiente sección destaca los múltiples desafíos y falencias que debe enfrentar la moderación de contenidos en Colombia.

Falta de transparencia y de medidas procesales

Existe una falta ampliamente reportada de transparencia en torno a las reglas que se aplican a la moderación de contenido y cómo se implementan. Esto genera preocupación

en el contexto colombiano y es notorio en las declaraciones de las partes interesadas entrevistadas.

Aunque los entrevistados tienen visiones distintas de las plataformas de redes sociales, en general comparten la preocupación por la falta de transparencia tanto en la comprensión de los términos de servicio de las plataformas como en las prácticas de moderación de contenido. Los entrevistados indicaron que los usuarios saben que cuando se inscriben a una plataforma de redes sociales deben firmar un acuerdo que es largo y difícil de entender, por lo que la mayoría no lo lee. Existe la creencia generalizada de que los procesos de moderación de contenidos en las plataformas son poco claros y sesgados. Aunque algunas plataformas han hecho notables esfuerzos por ofrecer más información y transparencia en este asunto (las normas comunitarias de las plataformas de redes sociales más usadas en Colombia se encuentran en español), todavía se quedan cortas en permitirle a los usuarios entender por completo las políticas y decisiones de moderación de contenido.[18]

Las partes interesadas entrevistadas destacaron que no era claro cómo las normas comunitarias se aplicaban a distintas categorías de contenidos potencialmente 'perjudiciales' y cómo eran moderadas por las plataformas.[19] Lo que más complica el asunto desde la perspectiva de los usuarios es que deben tener en cuenta múltiples políticas dentro de cada plataforma para entender qué tipo de contenido está permitido en cada una.

En cuanto a cómo se aplican las políticas, parece que ninguna plataforma está preparada para proporcionar el nivel de información sobre las prácticas de moderación de contenidos en Colombia necesario para ayudar a los usuarios a obtener una comprensión significativa de la aplicación local de las normas comunitarias. Tampoco están dispuestas a proporcionar los datos que ayudarían a seguir y servir como mecanismo de supervisión de las peticiones del gobierno. La ausencia de información a nivel de país, de datos clasificados según el volumen de contenido eliminado, las razones en las que se basó la eliminación, el tipo de moderación, el origen de la moderación, el número de apelaciones

recibidas y su estado, y las normas locales relacionadas con las reglas de moderación, entre otros indicadores, es una gran barrera para la incidencia a nivel local.

La representante de ColombiaCheck que entrevistamos, una de las dos organizaciones colombianas de *fact-checking* que firmaron el International Fact-Checking Network (IFCN) en Poynter, indicó que el primer paso que las plataformas deberían tomar es ofrecer más información sobre sus políticas, más allá del acuerdo inicial que aceptan los usuarios. Indicó también que 'cuando las plataformas empiezan a invertir en métodos de enseñanza, tienen que empezar a transferir el conocimiento de sus propias políticas y cuáles son los contenidos que clasifican como xenófobos o que pueden motivar nuevos ataques'.

La entrevistada de Artemisas, una organización feminista, indicó que 'el proceso no debería permitir que se elimine un tuit, por ejemplo, sin que se explique por qué y cómo se tomó esa decisión, sin enseñar. Creo que muchas de las normas de las plataformas no son comprensibles para las personas'.

Incluso si las plataformas han hecho esfuerzos razonables por ofrecer información y explicaciones, esto no es suficiente para explicar el impacto local que tienen sus prácticas de moderación y curación de contenido. Por ejemplo, mientras las empresas de redes sociales mencionan que cumplen con la legislación de los países en donde operan, no es claro hasta qué punto la legislación nacional se aplica o tiene impacto en sus decisiones, incluso en Colombia.

La falta de transparencia también impacta la capacidad para recurrir a soluciones frente a las acciones de moderación de contenido. Algunas de las partes interesadas entrevistadas expresaron su preocupación sobre problemas más estructurales y la ausencia de recursos legales para oponerse a las decisiones de moderación. Por ejemplo, la vocera de El Veinte, una organización que trabaja en libertad de expresión, indicó que 'en moderación de contenidos algunos aspectos imitan los sistemas de justicia informal, pero no tienen elementos básicos y garantías de debido proceso que dichos sistemas le

ofrecen a las partes. Creo que es necesario establecer estándares claros, eficientes y concertados'.

Por su parte, el representante del centro de estudios Linterna Verde comentó que, en la práctica, debido a la cantidad de contenido en línea que circula en redes sociales, las plataformas digitales no tienen la capacidad de cumplir con sus propios términos de servicio. Él recordó un caso muy sonado de moderación que actualmente se encuentra ante la Corte Constitucional:

> *El caso de Esperanza Gómez resume bien el asunto en cuestión: el contenido que ella subió a Instagram causó que se eliminara su cuenta sin recibir ninguna explicación de la regla que infringió. Ella tampoco recibió respuesta alguna ante su apelación incluso después de enviar varios correos electrónicos haciendo seguimiento. Esto llevó a que ella creara una nueva cuenta y tuviera que construir nuevamente su audiencia desde cero.*

Otra preocupación más reciente se debe a las prácticas de curación de contenido, que los entrevistados perciben como menos claras y cercanas a la censura. Los desafíos en términos de transparencia, y por lo tanto la capacidad efectiva para recurrir a soluciones, son especialmente pronunciados cuando se trata de la curación de contenido o las medidas de moderación de contenido que son menos evidentes que la eliminación de este o la suspensión de la fuente. Por ejemplo, la restricción de acceso al servicio, la desmonetización o el poner avisos de advertencia sobre cierto contenido.

El Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) de Argentina y de la región mencionó la falta de comprensión de los usuarios sobre la disminución de la visibilidad (*downranking)* y otras prácticas de curación y moderación. En la entrevista afirmó que 'se trata de prácticas muy opacas que se mueven fuera del radar y sobre las que la sociedad civil o el mundo académico tienen muy poca información. Además, estas prácticas son totalmente discrecionales. Las razones que desencadenan este tipo de prácticas no se comunican y no son claras'.

La organización uruguaya y regional de defensa de la libertad de expresión Observacom argumentó que la disminución de la visibilidad (*downranking)*, en la práctica, tiene los mismos efectos que la eliminación de contenidos porque impide a los usuarios ver el contenido y no está sujeto a la notificación al usuario mediante el debido proceso o reparación. En su opinión, 'la moderación de contenidos se ha estudiado a fondo, pero no ha ocurrido lo mismo con la curación de contenidos'.[20]

Del mismo modo, los miembros de la sociedad civil relacionan algunas de sus experiencias problemáticas en redes sociales con consecuencias de la moderación de contenidos difíciles de detectar.[21] La organización de derechos civiles, Temblores, por ejemplo, se refirió a la práctica del *shadowbanning*, en la que se oculta o se reduce la visibilidad de los contenidos de los usuarios sin que la plataforma les informe. Esta sensación de censura también se mencionó en el informe Guns versus Cellphones de Karisma, que estudió la protesta social en 2021.[22]

Las entrevistas evidenciaron la necesidad de explicar mejor cómo los contenidos que pueden infringir las normas de las empresas son objeto de distintos tipos de medidas o de curaduría, puesto que la retirada de contenidos ya no es la única sanción. El Veinte afirmó que 'también necesitamos más claridad sobre otras formas mediante las cuales las plataformas moderan el contenido, porque suprimir o amplificar no son las únicas formas en que los algoritmos permiten ocultar ciertas expresiones; a veces no es que la plataforma haya eliminado la publicación, sino que simplemente se la ha apartado de la vista'.

A las partes interesadas les gustaría ver un debate más abierto, que incluyese representantes de las plataformas de redes sociales, sobre la moderación de contenidos para propiciar la rendición de cuentas. Según la Fundación Gabo, organización para el desarrollo de los medios de comunicación:

> *Creo que esto debería convertirse en una cuestión de debate público y que deberíamos encontrar la manera no sólo de que haya transparencia, sino también de que los sistemas rindan cuentas y se gestionen de forma responsable para el público, porque*

están trabajando con bienes públicos. Que no es sólo la información; es también la vida privada de las personas. Es la paz. Entonces sí creo que más allá de que hayan desarrollado la tecnología y que traten de ser responsables, definitivamente es un tema que nos involucra a todos.

El entrevistado del medio de comunicación Liga contra el Silencio dijo:

Creo que la moderación debería debatirse y debería ser un tema más público que tenemos que entender mejor. ¿Cuáles son las normas? Eso es algo de lo que ni la audiencia ni los creadores de contenidos tienen ni idea. Aquí estamos adivinando cosas, por ensayo y error, pero debería ser muy público, la moderación debería ser un acuerdo muy claro en el que las fundaciones de libertad de expresión tuvieran algo que decir, no una decisión absolutamente unilateral.

Del mismo modo, La Silla Vacía indica que 'desde el periodismo, creo que sería muy útil formar parte de la discusión para intervenir en la opacidad que existe en torno a la moderación de contenidos. Desde ahí también podemos luchar contra la desinformación'.

Finalmente, vale la pena mencionar que las entrevistas con miembros de la sociedad civil y de los medios de comunicación mostraron que un tema relevante para los interesados en Colombia es la transparencia de los contenidos publicitados frente a los contenidos orgánicos. En otras palabras, cómo se puede reconocer el contenido pago -que los usuarios han pagado a las plataformas para publicar- en comparación con el contenido que ha sido publicado por los usuarios.

Evidentemente, es necesario aumentar las obligaciones de transparencia. Esto afecta no sólo a las mismas plataformas, sino también a los actores estatales que utilizan diferentes canales para influir en la moderación de contenidos. Como se explicará en la sección relacionada con las facultades legales estatales para solicitar la moderación de contenidos, las obligaciones de transparencia no se han desarrollado en Colombia, ni siquiera en el espacio regulado localmente de los operadores de telecomunicaciones, donde la facultad legal de bloquear contenidos existe desde hace más de 20 años.

Deficiencias de la moderación de contenidos automatizada

Sólo es posible que las plataformas moderen contenidos a escala si dependen, en cierta medida, de la moderación de contenidos automatizada, ya que la moderación humana sería incapaz de procesar la cantidad de información generada por los usuarios. Aunque la revisión humana es esencial para la interpretación de contenidos específicos en los contextos de sensibilidad cultural, creencias o sistemas de valores; la supervisión de contenidos en línea en tiempo real es una tarea gigantesca que puede no ser del todo factible sin la asistencia de la tecnología. Estos nuevos retos incluyen la preocupación por la situación de los moderadores de contenidos en Colombia. Al mismo tiempo, estas herramientas pueden suponer un grave riesgo para la libertad de expresión.

Desde la perspectiva del usuario, la moderación de contenidos automatizada es claramente un problema emergente y está vinculada a una serie de retos, como la falta de transparencia en torno a la moderación de contenidos, la ausencia de decisiones con enfoque lingüístico y cultural, y el hecho de obstaculizar a los medios de comunicación y a la información de interés público mediante la remoción de contenidos relacionados con grupos extremistas o violaciones de los derechos humanos.

Para los actores entrevistados, un elemento clave para entender la moderación de contenidos automatizada es la información que proporcionan las plataformas sobre el proceso de automatización y cómo se lleva a cabo en Colombia. Sin embargo, persiste una notable falta de transparencia sobre el grado de empleo de herramientas automatizadas en la moderación de contenidos. Según Meta, el 90% de lo que se identifica como contenido problemático se modera a través de sistemas automatizados. Meta informa sobre el índice de detección proactiva de contenidos, es decir, los contenidos o cuentas sobre los que se actúa para aplicar decisiones de moderación de contenidos antes de que los usuarios los denuncien. Sin embargo, no se encuentran datos similares para otras plataformas.

El portavoz de La Silla Vacía, la otra organización de verificación de hechos de la IFCN, mencionó:

> *Creo que los procesos de moderación de contenidos no son claros, no queda claro quién toma las decisiones de moderación, y hasta qué punto éstas se rigen únicamente por decisiones algorítmicas, o el impacto que la moderación tiene en los moderadores. Además, creo que las plataformas de redes sociales deberían seguir esforzándose por hacer más claras sus normas sobre contenidos permitidos y prohibidos.*

Moderación de contenidos y periodismo de interés público

Los periodistas colombianos son conscientes de la importancia de las plataformas digitales para la difusión de información, sobre todo fuera de las grandes ciudades. Como resultado, los medios de comunicación han aumentado su presencia digital y han empezado a difundir contenidos en Facebook, Instagram, TikTok, X y otras plataformas.

Sin embargo, depender de las plataformas de redes sociales para la difusión puede tener repercusiones negativas. Las deficiencias de los sistemas automatizados de moderación de contenidos a la hora de impedir la circulación de contenidos 'perjudiciales' han complicado las cosas en ocasiones para los medios de comunicación. Como explica la Liga contra el Silencio:

> *Hay un valor periodístico muy importante, que es llamar a las cosas como son, decir feminicidio, decir homicidio, decir reinsertado (excombatiente). Pero en las plataformas hemos tenido que silenciar esas palabras en los vídeos o escribirlas con números para que el algoritmo no 'castigue' el contenido. Hemos tenido que usar eufemismos que también acaban tendiendo un manto de duda sobre algunas luchas. Eso ha sido difícil.*

Esta es una queja seria que debería ser atendida. Hay temas importantes para el debate público que, para tener presencia en el espacio público digital, acaban pasando por mecanismos de encubrimiento que les permiten evitar la moderación de contenidos. Varias personas que cubrieron las protestas de 2021 denunciaron una situación similar (véase el estudio de caso 1).

Las estrategias que la prensa debe implementar para asegurar que sus contenidos permanezcan en línea, a pesar de su interés público, evocan las discusiones que el Ministerio de Cultura describió como un agujero negro digital.[23] Uno de los ejemplos que la entonces Ministra de Cultura dio para ilustrar este agujero negro fue la desaparición de las redes sociales de las voces de la guerrilla de las FARC durante el proceso de paz.

El Ministerio también señaló otra cuestión que puede impedir informar e investigar sobre asuntos de interés público: las listas de grupos extremistas prohibidos de las plataformas. El Ministerio explicó que para cualquier persona interesada en el proceso y el acuerdo de paz de Colombia de 2016, las cuentas de redes sociales de sus protagonistas serían una importante y singular fuente primaria de información. Sin embargo, la ministra advirtió que al estar las FARC en la lista de grupos terroristas internacionales, las plataformas de redes sociales -como X, Facebook o YouTube- bloquean con frecuencia contenidos y cancelan las cuentas de las FARC, a pesar de ser parte del proceso de paz que se desarrolla con el Gobierno.

Estudio de caso 1: Protestas sociales en 2021[24]

Durante las crisis sociales, las plataformas se enfrentan a importantes retos en sus procesos de toma de decisiones sobre moderación de contenidos, curación de contenidos y mecanismos de apelación. En estos casos, la falta de transparencia descrita y los efectos de la automatización tienden a empeorar la situación. El volumen de información que se produce durante los momentos de agitación social, la concentración de la producción y publicación de contenidos por parte de usuarios individuales en determinados momentos -y su naturaleza (por ejemplo, los contenidos que denuncian abusos policiales pueden clasificarse como contenidos violentos prohibidos por las normas de la comunidad)- suponen una tensión adicional para los procesos de moderación de contenidos durante las protestas. Además, las personas

vinculadas a esos acontecimientos se muestran especialmente críticas con las prácticas de moderación de contenidos que perciben como injustas o mal motivadas.

Durante las protestas sociales de 2021 en Colombia, los sistemas de moderación, curación y apelación de contenidos de las plataformas aparentemente no funcionaron de forma adecuada, lo que afectó a las personas que estaban vinculadas a las protestas o que proporcionaban información sobre las mismas.

El 28 de abril de 2021, después de que el presidente Iván Duque presentara un proyecto de reforma tributaria ante el Congreso, se inició en Colombia un movimiento colectivo de protesta ciudadana conocido como 'Paro Nacional'. Desde entonces y hasta el 15 de junio de 2021, se convocaron marchas y actividades en varias ciudades del país.

En medio del malestar social y democrático y de los enfrentamientos con las autoridades, se hicieron virales las publicaciones en redes sociales que documentaban el uso excesivo de la fuerza por parte de las fuerzas públicas contra los ciudadanos o los ataques de éstos contra agentes e infraestructura local.

También se emitieron informes sobre posibles acciones del Estado para limitar los derechos a la libertad de expresión y reunión, el acceso a la información y la privacidad de los ciudadanos, ejercidos en o a través de plataformas de redes sociales.

Entre las medidas adoptadas por el Estado estuvo el ciberpatrullaje de redes sociales. Esto permitió al Estado elaborar perfiles de personas y contrarrestar el discurso que se desviaba de la narrativa del gobierno. Otras medidas incluyen la vulneración de la libertad de expresión, de asociación y de reunión por no brindar información sobre las interrupciones al acceso a Internet en lugares de alta concentración, la búsqueda y revisión de contenidos en los dispositivos móviles de los manifestantes, a menudo sin

su consentimiento, y la orden a los proveedores de servicios de Internet de bloquear las páginas que contuvieran información sobre miembros de la fuerza pública

Jahfrann, fotógrafo freelance colombiano residente en Cali,[25] relató la tensión de documentar esta situación en redes sociales durante esos días. Muchas personas denunciaban inexplicables cortes puntuales de Internet que afectaban a su capacidad para usar las redes sociales. Dijo:

> *Digamos que en este momento se ha calmado porque ya hay un pronóstico nacional e internacional, pero cuando apenas empezó, parecía que toda la red estaba perdida. Siloé -un barrio de Cali- estuvo apagado un día durante cinco horas, cinco horas en las que no entró ni salió nada del barrio. Quiero decir, yo hablaba por walkie-talkie con la gente de derechos humanos, y no sabíamos si estaban vivos o muertos -simplemente no había señal-, vi la unidad móvil con grandes equipos y una antena. No tengo foto. Gran parte de lo que compartí fue 'atemporal' porque allí no había señal.*

La Fundación para la Libertad de Prensa (FLIP) denunció que el 6 de mayo la cuenta X de Noís Radio, un medio independiente de Cali (@noisradio), había sido 'reiteradamente restringida'.

Noís Radio interpuso recurso contra la decisión de bloqueo, pero la restricción se impuso en varias ocasiones, lo que le llevó a manifestar que los procedimientos de recurso en las redes sociales eran ineficaces. El medio de comunicación sostuvo que las etiquetas impuestas a su cuenta acabaron silenciando su voz, por cuanto los avisos de advertencia eran visibles para los usuarios.[26] Las restricciones y bloqueos de cuentas llevaron a algunos periodistas a recurrir al código alfanumérico para esquivar al algoritmo.

Los contenidos publicados en momentos de agitación social son de suma importancia porque sirven para documentar violaciones de derechos humanos. Alejandro Gómez, de la Liga Contra el Silencio, trabajaba entonces en el portal digital 070, un medio que reconstruía sucesos violentos durante las protestas, incluido el asesinato del manifestante Dilan Cruz en 2019. Señaló que 'no tenían capacidad para cubrir la violencia policial', por lo que era imposible verificar o cuestionar las versiones oficiales de los hechos.

Aunque se identificaron casos de restricción de expresiones en plataformas de redes sociales durante las protestas de 2021, Meta se limitó a explicar sus problemas de software, mientras que X no explicó en absoluto su falla. La decisión de Meta de admitir públicamente el problema de software y sus consecuencias es una buena práctica que deberían adoptar otras plataformas en casos similares. Sin embargo, las explicaciones de Meta no estaban relacionadas con los problemas de moderación de contenidos y los mecanismos de mitigación durante las protestas.

Además, periodistas y bibliotecarios creen que la moderación de contenidos afecta al panorama informativo. Algunas situaciones producen un efecto de supresión excesiva que repercute en la sociedad en general, debido a la enorme dependencia de los medios digitales y las fuentes de memoria de la información en las redes sociales.

Contexto en los procesos de moderación y curación de contenido

Hay una serie de informes de investigación y publicaciones de medios de comunicación sobre la moderación y la curación de contenidos en Colombia. Algunos ejemplos de estos informes son: el informe sobre moderación de contenidos con derechos de autor de Karisma, el análisis de casos de Linterna Verde y el informe sobre moderación y libertad de prensa de Observacom. Estos informes confirman que las decisiones de moderación de contenidos deben interpretarse en el contexto local y tener en cuenta las diversas

particularidades jurídicas, políticas, culturales y lingüísticas, o la protección especial existente de determinadas poblaciones a nivel local.

Cuando planteamos preguntas sobre el contexto local a los entrevistados representantes de plataformas de redes sociales durante esta investigación, Meta mencionó que incorporan diferentes voces y puntos de vista a la hora de redactar las políticas. X hizo hincapié en el aspecto global de la conversación, afirmando que los equipos que trabajan en los mecanismos de detección están formados para tener en cuenta la diversidad y el contexto. En este sentido, Google afirmó que la diversidad se tiene en cuenta en la elaboración de sus políticas.

Meta y Google tienen oficinas locales en Colombia que no se limitan a labores de mercadeo, sino que cuentan con funcionarios encargados de las políticas y mantienen buenas relaciones con la sociedad civil. X solía tener un equipo ubicado en Ciudad de México que se ocupaba de cuestiones políticas con la sociedad civil latinoamericana, pero se desmanteló en 2023. No hay información sobre los equipos que moderan los contenidos locales ni dónde están ubicados. Hasta ahora, estas empresas han desarrollado canales directos con algunas organizaciones locales de la sociedad civil en circunstancias especiales (por ejemplo, durante la protesta de 2021 o las elecciones de 2022). Los canales se activan en esos momentos y las empresas han mencionado que toman medidas especiales que incluyen la moderación de contenidos local con matices, como se describe más adelante en este documento. Un periodista reveló que había moderadores de TikTok con sede en Colombia; sin embargo, no es claro qué papel desempeñan en la moderación de contenidos a nivel local.

Además, algunas plataformas con presencia en Colombia revisan estudios de caso al intentar abordar el reto del contexto. Esto lo hace el Consejo Asesor de Contenidos de Meta y también una iniciativa de YouTube que selecciona casos de interés para el contexto nacional. Sin embargo, a pesar de la reciente pandemia y los graves conflictos sociales en Colombia, estas herramientas sólo se han utilizado una vez. Por ejemplo, el Consejo Asesor de Contenidos de Meta seleccionó y decidió sobre el caso del uso de la

palabra 'marica' (véase el estudio de caso 2). YouTube, por su parte, relató cómo decidió no eliminar de su plataforma un documental crítico sobre el ex presidente Álvaro Uribe. Estos esfuerzos de las plataformas no son suficientes en términos de transparencia; y son también insuficientes a la hora de proporcionar información para que las plataformas y los usuarios actúen en consecuencia.

Los entrevistados también establecieron una conexión entre los sistemas automatizados y la ausencia de decisiones con matices lingüísticos y culturales. Una representante de Plurales, un *think tank* de la Universidad del Rosario, dijo: 'Es una tarea que no puede hacer una máquina. Por ejemplo, la palabra 'marica' tiene muchos significados: puede ser un cariño, pero también una palabra despectiva, no necesariamente homofóbica, pero puede tener usos homofóbicos y transfóbicos. Las máquinas no pueden entender estas diferencias'. La entrevistada señaló el citado caso colombiano que fue seleccionado por el Consejo Asesor (véase el estudio de caso 2), un caso en el que el presidente de entonces fue llamado 'marica' durante las protestas de 2021.

Linterna Verde añadió sobre este tema: 'cuando la moderación se convierte en una serie de palabras prohibidas... sin un conocimiento suficiente del contexto y sin evaluar la situación en detalle, conduce a errores', apuntando así al problema del contexto.

Para Meta, este caso demuestra que es posible corregir un error cometido en la moderación de contenidos teniendo en cuenta el contexto local. Meta declaró en la entrevista que 'la decisión del Consejo Asesor relacionada con la moderación de contenidos de las publicaciones que utilizaron la palabra 'marica' en Colombia se tendrá en cuenta en futuros casos. La moderación de contenidos no es estática, evoluciona. Si se encuentran excesos en la interpretación de una norma de moderación de contenidos, la plataforma puede limitarlos'. En línea con esto, Meta afirmó que 'las actividades de moderación son tan dinámicas como el propio contenido. Si las sociedades exploran nuevas tendencias y movimientos sociales, el panorama de la moderación de contenidos cambia continuamente. Siempre está cambiando'.

Esta no es la única forma en que las plataformas contextualizan la moderación de contenidos. En cuanto a la aplicación contextual de las normas comunitarias, existe una advertencia para los períodos de agitación social. Las plataformas afirmaron que tienen procedimientos especiales durante las elecciones y que se adoptaron enfoques similares en otros momentos de conflicto social. X informó a las investigadoras que la adopción de medidas en épocas normales no es la misma que en épocas atípicas. Durante las crisis humanitarias, las pandemias o los periodos electorales, X adopta una política de relativa a la información engañosa en situaciones de crisis.[27]

Meta señaló que gran parte de su trabajo consiste en anticiparse a acontecimientos políticos o periodos de incertidumbre política. Pueden establecer temporalmente un Centro de Integridad de Operaciones de Producto, 'que es un grupo de trabajo compuesto por expertos en la materia de nuestros equipos de producto, política y operaciones, [que] permite a estos expertos detectar, clasificar, investigar y mitigar más rápidamente los riesgos en la plataforma', según la Actualización Trimestral del Consejo Asesor. En estos momentos, Meta también recibe asesoramiento de socios locales (a través de su programa de socios de confianza (*Trusted Partner)*) que entienden las especificidades de cada contexto, y establece conversaciones con las comisiones electorales.[28] Por último, Meta también mencionó la existencia de un programa de *fact-checking* por terceros durante las elecciones.

Otra cuestión relacionada con el contexto en la moderación de contenidos es su impacto desproporcionado en determinados grupos en comparación con otros, como lo describió Linterna Verde:

> *Las plataformas basan sus normas sobre discurso de odio en categorías internacionalmente protegidas como la nacionalidad, la orientación sexual o la raza. Sin embargo, se olvidan de proteger otras categorías que están protegidas en el contexto del conflicto armado colombiano: en particular, los defensores de los derechos humanos, los excombatientes o los periodistas que, entre otros, hablan sobre el conflicto y, por lo tanto, están mucho más expuestos a ataques directos.*

En un informe sobre la violencia de género contra mujeres periodistas en Colombia, las periodistas mencionaron que no utilizaban los mecanismos de respuesta disponibles en las plataformas debido a la falta de conocimiento sobre su existencia y funcionamiento o porque creen que dichos mecanismos son ineficaces. El informe recomendaba a las empresas 'comunicar de manera más efectiva, accesible y en los idiomas locales los mecanismos de respuesta disponibles para atender las violencias basadas en género que ocurre en sus plataformas'. También recomendaba que las empresas realizaran consultas periódicas para mejorar sus políticas y prácticas.

La moderación de contenidos cuando se trata de figuras de reconocimiento público local también es motivo de preocupación para los actores entrevistados. Tanto el caso seleccionado por el Consejo Asesor (véase el estudio de caso 2) como el caso seleccionado por YouTube sobre la petición de retirar un documental sobre el expresidente Álvaro Uribe hablan de cómo las figuras públicas (el presidente colombiano en el caso de Facebook o un partido político y expresidente en el caso de YouTube) deben tener una mayor tolerancia al escrutinio público y al cuestionamiento por parte de las audiencias, y también una mayor responsabilidad cuando crean contenidos para su distribución.

Este aspecto está relacionado con las reflexiones de algunos de los entrevistados. La forma en que la moderación de contenidos aborda la responsabilidad de las figuras públicas es una preocupación para los actores locales. *Sentiido*, un medio de comunicación con enfoque LGBTQI+ (lesbianas, gays, bisexuales, transexuales, queer e intersexuales), afirmó: 'Me parece que falta un debate sobre la responsabilidad de la visibilidad. Es decir, una persona cuya visibilidad en los medios de comunicación es tan importante tiene una responsabilidad en la promoción de discursos no violentos contra comunidades que han sido históricamente marginadas'.

Estudio de caso 2: insulto 'marica' durante protestas sociales

El Relator Especial de las Naciones Unidas sobre la promoción y protección del derecho a la libertad de opinión y de expresión publicó un informe sobre la moderación de contenidos en línea generados por usuarios en el que describía una paradoja a la que se enfrenta la moderación de contenidos en plataformas: aunque 'las empresas hacen hincapié en la importancia del contexto a la hora de valorar la aplicación de las restricciones de carácter general... La revisión detallada del contexto puede verse frustrada por las limitaciones de tiempo y de recursos que adolecen los moderadores humanos, la dependencia excesiva de la automatización o la insuficiente comprensión de los matices lingüísticos y culturales'. La moderación de contenidos es un ecosistema complejo de tratar, y el caso de moderación de contenidos en Colombia seleccionado por el Consejo Asesor de Meta y decidido en octubre de 2021 describe esta situación:

> *En mayo de 2021, la página de Facebook de un medio de comunicación regional colombiano compartió una publicación de otra página de Facebook sin añadir ninguna descripción adicional. El contenido en cuestión en este caso es dicho post compartido. La publicación original contiene un breve vídeo que muestra una protesta en Colombia con personas marchando detrás de una pancarta que dice 'SOS COLOMBIA'.*
>
> *Los manifestantes cantan en español y se dirigen al presidente colombiano, mencionando la reforma tributaria recientemente propuesta por el gobierno. Como parte de su cántico, los manifestantes llaman al presidente 'hijo de puta' una vez y dicen 'deja de hacerte el marica en la tv' una vez. Facebook tradujo estas frases como 'son of a bitch' y 'stop being the fag on tv'. El video va acompañado de un texto en español en el que se expresa admiración por los manifestantes. El post compartido tuvo unas 19.000 visitas, y menos de cinco usuarios lo reportaron a Facebook.*

> *Facebook eliminó el contenido porque contenía la palabra 'marica' Esto infringía la Norma Comunitaria sobre lenguaje que incita al odio de Facebook, que no permite contenidos que 'describe o se dirige de forma negativa a las personas con insultos' basadas en características protegidas como la orientación sexual. Facebook señaló que, aunque en teoría podría aplicarse a este tipo de contenidos el* criterio de *interés periodístico, este criterio sólo puede aplicarse si los moderadores de contenidos que revisan inicialmente el contenido deciden elevarlo para que sea revisado por el equipo de política de contenidos de Facebook. Esto no ocurrió en este caso. (resumen del caso por el Consejo Asesor de Meta)*

El Consejo Asesor revocó la decisión de Facebook de retirar la publicación. El Consejo Asesor concluyó que, aunque la retirada del contenido por parte de Facebook parecía ajustarse a su Norma Comunitaria sobre Expresión de Odio, debería haberse utilizado el criterio de interés periodístico para que el contenido siguiera en línea. Según las observaciones del público y las recomendaciones de los expertos, la palabra 'marica' tiene varias connotaciones y podría utilizarse sin tener intención discriminatoria. Los expertos explicaron que el término había alcanzado un uso generalizado en Colombia para referirse a una persona como 'amigo' o 'amigo', y también como insulto como 'estúpido', 'tonto' o 'idiota'. Sin embargo, hubo consenso en que sus orígenes eran homofóbicos y que se utilizaba especialmente contra hombres homosexuales. El Consejo Asesor señaló:

> *El criterio de interés periodístico requiere que Facebook evalúe el interés público de admitir una determinada expresión frente al riesgo de daño derivado de admitir contenidos infractores. Para ello, Facebook tiene en cuenta la naturaleza de la expresión, así como el contexto específico del país, su estructura política y si goza de libertad de prensa.*

Al evaluar el valor de interés público de este contenido, el Consejo observa que se publicó durante las protestas masivas contra el gobierno colombiano en un momento significativo de la historia política del país. Aunque los participantes parecen utilizar el término difamatorio deliberadamente, se utiliza una vez entre otras muchas expresiones y el cántico se centra principalmente en la crítica al presidente del país.

El Consejo también apunta que, en un entorno en el que las vías de expresión política son limitadas, las redes sociales han proporcionado una plataforma para que todas las personas, incluidos los periodistas, compartan información sobre las protestas. Aplicar el criterio de interés periodístico en este caso significa que sólo se permitiría un contenido 'perjudicial' excepcional y limitado.

Facultades legales estatales para solicitar moderación de contenidos

No existe una regulación específica que rija integralmente la moderación de contenidos en Colombia. Sin embargo, existen regulaciones dispersas que exigen la implementación de órdenes de bloqueo a los operadores de telecomunicaciones. Estas regulaciones no son sólo para el Material de Abuso Sexual Infantil ('CSAM' por sus siglas en inglés);[29] también existen para hacer frente a los juegos de azar[30] e incluyen todas las demás órdenes legales provenientes de las autoridades administrativas.[31] Además, puede haber órdenes judiciales y administrativas de bloqueo -como medidas cautelares en una acción de tutela- que tengan lugar durante estados de emergencia y excepción.[32] Todas estas órdenes emitidas a los operadores de telecomunicaciones se canalizan a través del MinTIC.

De acuerdo con los estándares internacionales y constitucionales, cualquiera que sea su base legal o las autoridades responsables, cualquier orden de bloqueo o restricción debe cumplir con el triple criterio de legalidad, objetivo legítimo, necesidad y proporcionalidad. Sin embargo, aunque no existe un análisis legal detallado de las regulaciones dispersas antes mencionadas desde la perspectiva de la libertad de expresión en Colombia, es preocupante que algunas órdenes de bloqueo hayan sido tan amplias como para bloquear casi por completo el acceso a páginas web enteras. Este fue el caso de RapidShare para los usuarios de Telefónica en 2010, y fue caso similar para todos los usuarios en Colombia con InternetArchive en 2021 durante la protesta social. La página web de InternetArchive terminó siendo bloqueada sólo por Avantel y Emcali porque las otras compañías no acataron la orden después de que su propio análisis demostrara que era desproporcionada.

Por otra parte, las plataformas de redes sociales informan sobre solicitudes para retirar contenido por motivos legales en sus informes de transparencia. Esto es algo que los representantes de Google, X y Facebook confirmaron en las entrevistas. El problema es que esos informes no tienen suficiente información para analizar qué marcos normativos están aplicando, o si sus normas de moderación de contenidos contradicen las normas

locales aplicables. Esto se complica cuando el informe de transparencia de X agrega las cifras de aplicación de sus normas comunitarias a las solicitudes gubernamentales.

La regulación vigente no obliga a las plataformas de redes sociales a ser plenamente transparentes sobre las prácticas de moderación de contenidos ni exige al gobierno colombiano facilitar información sobre las peticiones (como las solicitudes de eliminación) elevadas a las plataformas de redes sociales. Cuando las investigadoras preguntaron sobre la regulación de la moderación y la curación de contenidos, el MinTIC les informó:

> *La regulación de proveedores de redes y servicios de telecomunicaciones requiere bloquear los sitios web con contenidos prohibidos de las listas con URLs que contienen pornografía infantil[33] emitidas por la Dirección de Investigación Criminal e Interpol (DIJIN), publicadas en la página web del MinTIC.*

El MinTIC añadió que 'el legislador no atribuyó a esta entidad la competencia para regular, vigilar y controlar la provisión de contenidos y aplicaciones o plataformas tecnológicas'. Aunque esto sea cierto, como ya se ha descrito, las órdenes de bloqueo de contenidos en la regulación actual de las telecomunicaciones no son sólo para el CSAM. En todo caso, el MinTIC no menciona su papel más amplio en cuanto a sus diversas capacidades de bloqueo de contenidos. La declaración escrita también dice que el MinTIC entiende que cualquier regulación de la información o los contenidos en las redes sociales debe abordarse con la máxima cautela, ya que podría implicar una limitación sobre derechos fundamentales.

Aunque todavía no existe regulación de las plataformas, es posible que esto cambie pronto. Llama la atención que la gran mayoría de las actuales propuestas de regulación estatal de redes sociales se refieran al control de contenidos lícitos – contenidos que, *prima facie*, están protegidos por la libertad de expresión.

Recientes proyectos de ley han tratado de imponer mecanismos de filtrado y bloqueo para proteger a los menores de edad de contenidos potencialmente 'perjudiciales', o prohibir

cualquier discurso sobre prostitución o promoción de actividades sexuales en las redes sociales que se desvíe de la visión que los reguladores tienen del fenómeno. Algunos proyectos de ley pretenden imponer funciones a las autoridades administrativas para que bloqueen con prontitud y rapidez los contenidos de pago que puedan afectar a los derechos políticos de las mujeres. Estos casos se alinean con las presiones regulatorias en América Latina en países como Argentina o Perú, lo que podría indicar una tendencia más amplia en la región.

El CELE mencionó que existe una tendencia global en la que 'muchos de los proyectos y leyes que se promulgan hoy en día, incluyendo aquellos que regulan procesos, implican cierta renegociación sobre los límites legítimos de la libertad de expresión en las plataformas'. Los proyectos de ley que se han presentado en el Congreso en Colombia comparten esta característica - la definición de lo que es contenido indebido es generalmente ambigua y a menudo implica funciones de vigilancia activa de las redes sociales por parte de las plataformas y el Estado. Por ejemplo, la ley de violencia contra las mujeres en la vida política está siendo debatida y, al momento de redactar este informe, el proyecto está siendo evaluado por la Corte Constitucional porque existen diferentes puntos de vista sobre si sus disposiciones de moderación de contenidos cumplen con el marco constitucional colombiano. Por ejemplo, la ley de violencia contra las mujeres en la vida política está siendo debatida y, al momento de redactar este informe, el proyecto está siendo evaluado por la Corte Constitucional porque existen diferentes puntos de vista sobre si sus disposiciones de moderación de contenidos cumplen con el marco constitucional colombiano.

Una representante de Red PaPaz, una organización para la protección de la infancia que participa en el proceso de bloqueo del CSAM con los operadores de telecomunicaciones en Colombia, insistió en que les gustaría ver más acciones a nivel de las plataformas. La entrevistada dijo que 'es necesario entender que el abuso y la explotación de las imágenes de un niño se utilizan para la trata o para ganar dinero – se trata de un delito muy complejo. En estos casos, se han violado los derechos de un menor'.

Uso estatal de las normas comunitarias para restringir contenidos y perfiles

El mapeo de las regulaciones gubernamentales sobre moderación de contenidos de las plataformas en el informe de 2018 sobre moderación de contenidos en línea generados por usuarios incluía exigencias gubernamentales que no se basaban en leyes nacionales. El informe explicaba que estas demandas a menudo incluían presiones a las empresas para que aceleraran la retirada de contenidos mediante esfuerzos no vinculantes que habían evolucionado en acuerdos de coordinación entre empresas y Estados que pueden perjudicar la privacidad y la libertad de expresión.[34]

El Relator observó que los Estados se basan cada vez más en los términos de servicio de las plataformas de redes sociales para solicitar la retirada de contenidos que consideran censurables. Las prácticas de los Estados al solicitar la retirada de contenidos lícitos que pueden considerarse extremistas 'plantean la posibilidad de que los Estados puedan recurrir a unas condiciones de servicio de carácter privado para eludir las leyes o las normas de derechos humanos contrarias a las restricciones de contenido' (párr. 53).

No hay evidencia de que el gobierno colombiano tenga acuerdos formales con las plataformas para coordinar el monitoreo o remoción de contenidos. No obstante, las plataformas sí tienen formas de abordar los contextos locales de una manera más específica, lo que puede incluir la cooperación con autoridades públicas. Por ejemplo, durante las elecciones, las plataformas digitales pueden coordinarse formalmente con las autoridades electorales para informar a los votantes sobre las elecciones. También hay proyectos especiales, como el de YouTube Priority Flagger. Este programa incluye como socios a organismos gubernamentales y ONGs 'ya que son particularmente eficaces para notificarnos sobre el contenido que infringe los Lineamientos de la Comunidad'. Según la experiencia de Karisma - durante la protesta social y en periodos electorales más recientes -, las plataformas también ofrecen canales de denuncia especiales y más ágiles para que los socios proporcionen información específica.

Los informes de transparencia de algunas plataformas sí reconocen cómo el Estado colombiano utiliza las normas comunitarias para solicitar la retirada de contenidos;[35] sin

embargo, los informes sólo dejan entrever estas acciones. De nuevo, la ausencia de información dificulta la comprensión de los motivos de las solicitudes, el análisis de la solicitud según los estándares de derechos humanos en Colombia y la lista de autoridades responsables de dichas solicitudes.

Sin facultades legales, las autoridades colombianas habían estado enviando solicitudes a las plataformas utilizando las normas comunitarias. A partir de los datos proporcionados por algunas de las plataformas sobre Colombia,[36] durante la pandemia de Covid-19, el Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA) utilizó las normas comunitarias para controlar la circulación de la información. Durante las protestas de 2021, se utilizó el marco de la violencia y el terrorismo con el mismo fin.[37] El estudio de caso 1 aporta más datos sobre cómo se sintieron las personas durante las protestas de 2021 y cómo esto les impactó, ya que hablan de una sensación de censura e identifican al Estado como actor.

Al ser entrevistada, la representante del CELE advirtió de que se había subestimado el papel del Estado en este sentido. La entrevistada dijo que el Estado no es sólo un instigador de la moderación de contenidos que pide a las plataformas que retiren determinados contenidos o cuentas; también trata de resolver los problemas de la sociedad que no ha podido abordar mediante la moderación de contenidos. Los Estados no resuelven los problemas estructurales de la sociedad, como la discriminación o la violencia, por lo que las demandas de protección insatisfechas se trasladan a peticiones de bloqueo de contenidos o cuentas por parte de las plataformas. El CELE afirmó que las fallas de las instituciones públicas para abordar problemas complejos y la forma en que los funcionarios se han comportado en los debates públicos han llevado a que la 'esperanza de la gente [sobre la demanda insatisfecha en el discurso] se deposite en el mensajero, es decir, en el intermediario'.

Aunque las cifras comunicadas por las plataformas sobre el uso estatal de las normas comunitarias para controlar los contenidos son bajas en Colombia, la ausencia de razonamiento y publicidad detrás de las peticiones es problemática. El Veinte también

teme que las autoridades gubernamentales intenten reforzar cada vez más el control sobre la 'plaza pública digital' y exijan acciones de moderación de contenidos a las plataformas, con implicaciones negativas para la libertad de expresión.

Una coalición sobre moderación de contenidos y libertad de expresión

Esta investigación exploró la viabilidad de formar una coalición local sobre moderación de contenidos y libertad de expresión. Se encontró que:

1. En Colombia existen una serie de iniciativas que agrupan a organizaciones y movimientos que, ya sea directamente vinculados a los derechos digitales o formados en otras áreas de acción, pueden unirse para trabajar conjuntamente en temas concretos relacionados con la moderación de contenidos.
2. Colombia cuenta con organizaciones que trabajan en coalición y mantienen diálogos, ya sea a escala nacional o regional.
3. Hay puntos concretos en los que convergen los participantes en la investigación y otros agentes de múltiples partes interesadas, y en los que se puede trabajar para formar intereses comunes.
4. El gobierno y las plataformas de redes sociales son los actores más ausentes en este debate. Será necesario llevar a cabo una labor de promoción para exigir prácticas más respetuosas con los derechos en lo que respecta a la moderación y curación de contenidos. El enfoque más eficaz para lograrlo implicará que las partes interesadas locales encuentren formas de interactuar colectivamente con las plataformas.

Esta sección examina la propuesta de formar una nueva coalición en Colombia y cómo puede hacerse, teniendo en cuenta las especificidades de la sociedad civil colombiana y las coaliciones existentes que se ocupan de cuestiones de moderación de contenidos. También evalúa la propuesta de una red de organizaciones y coaliciones existentes como impulsor clave para establecer los puntos críticos que se llevarán a los actores relevantes en el proceso.

Formando una coalición potencial

Durante el transcurso de esta investigación, se mapearon los actores relevantes que trabajan en la intersección entre la moderación de contenidos en línea y la libertad de

expresión en Colombia. Se realizaron entrevistas (véase el anexo B) a diferentes partes interesadas para evaluar cómo entienden los actores locales la moderación de contenidos en las redes sociales. La investigación ha encontrado que actualmente no hay acuerdos entre los actores locales sobre lo que constituye un contenido potencialmente 'perjudicial' en las redes sociales y sobre las cuestiones clave que afectan a la moderación de contenidos. Una representante de Plurales mencionó:

> *Es muy difícil lograr una definición universal del concepto de perjuicio, por ejemplo, porque las organizaciones y los usuarios conservadores que ven información sobre orientaciones sexuales e identidades de género diversas pueden pensar que este contenido es potencialmente 'perjudicial' para sus hijos. Por eso es importante entender no sólo cómo consideraría yo que algo es perjudicial, sino también cómo está organizada la sociedad en términos de equilibrios de poder.*

Échele Cabeza, organización dedicada a la sensibilización sobre el consumo de drogas, afirmó:

> *Nosotros trabajamos con sustancias psicoactivas, que es un tema súper transgresor. Por ejemplo, podemos publicar un vídeo sobre cómo una persona puede inyectarse asumiendo menos riesgos. Para mí eso es información sanitaria, pero para otros es promoción del consumo de drogas, así que la percepción del riesgo y el peligro es algo que cambia para cada persona.*

Aunque las entrevistas mostraron una percepción general de que el perjuicio puede estar relacionado con expresiones que puedan suponer 'discurso de odio' y violencia, algunos entrevistados (véanse las dos citas anteriores) destacaron que ciertos temas son más complejos. Lo que algunos pueden percibir como información esencial que merece ser difundida, otros pueden considerarlo información perjudicial.

También es importante entender que actualmente en Colombia las discusiones sobre la moderación de contenidos surgen principalmente en el contexto de debates sobre temas específicos, como la violencia de género o elecciones. Estas experiencias deben tenerse

en cuenta en cualquier iniciativa que pretenda abordar la moderación de contenidos en Colombia.

Con este trasfondo, los elementos básicos para crear un grupo informado de partes interesadas que traten y defiendan las cuestiones clave relacionadas con la moderación de contenidos en Colombia son:

1. identificar temas o áreas concretas para trabajar en grupo;
2. identificar a los principales destinatarios de las campañas de sensibilización; y
3. armonizar la interpretación de las partes interesadas locales sobre la moderación de contenidos mediante el desarrollo de capacidades y los intercambios dentro del grupo.

Las siguientes secciones proporcionarán un análisis más detallado de estos requisitos, identificarán las necesidades de las distintas partes interesadas para llegar a un entendimiento común sobre la moderación de contenidos y presentarán algunas soluciones basadas en la realidad local colombiana.

Moderación de contenidos: un debate abierto para los actores locales

Actualmente, las organizaciones de la sociedad civil en Colombia reconocen en general que las experiencias de los usuarios en las redes sociales pueden ser diferentes y que hay ciertos grupos más afectados que otros por los contenidos 'perjudiciales' en línea, como el 'discurso de odio' o la 'desinformación'. En línea con la misión de sus organizaciones, algunos entrevistados mencionaron la necesidad de regular el discurso que es explícitamente racista, clasista, homofóbico, transfóbico, etc., mientras que otros expresaron su preocupación de que dicha regulación tenga el potencial de vulnerar la libertad de expresión.

Para precisar, algunas organizaciones consideran que ciertos tipos de contenidos discriminatorios deben eliminarse de inmediato de las redes sociales. Plurales afirma:

> *Es evidente que hay algunas palabras asociadas con el 'discurso de odio' y es importante ser razonable con lo que se dice. Así que, atendiendo al sentido común, si*

es obvio que un contenido profundiza en alguna desigualdad relacionada con cuestiones de género, clase, discapacidad, etnia y raza, creo que es importante eliminar ese contenido inmediatamente.

Por el contrario, una representante del CELE dijo que las definiciones de lo que constituye cada tipo de discurso que puede ser objeto de moderación de contenidos suelen ser vagas, no prevén excepciones cuando se trata de discursos protegidos, como el discurso político o de interés público, y pueden dar lugar a un control indebido sobre el discurso en línea. Además de afectar a la libertad de expresión, las prácticas de moderación de contenidos afectan gravemente al debate público.

Luisa Isaza, experta e investigadora colombiana en libertad de expresión y estudiante en la Universidad de Oxford, Reino Unido, resaltó la vulneración de la libertad de expresión que se produce cuando se eliminan contenidos que denuncian graves violaciones de derechos humanos, como es el caso de los falsos positivos, a pesar de no haber violado las normas de la comunidad. Otras investigaciones que analizan casos de contenidos que fueron moderados sin la debida justificación pueden aportar nuevos puntos de entrada al debate. Un investigador de la Universidad Externado, Colombia, validó esta afirmación diciendo que 'hay muchos casos límite o sospechosos cuya eliminación no está debidamente justificada que pueden colarse. En estos casos, es evidente que las normas y procedimientos de retirada de contenidos pueden interferir indebidamente en la libertad de expresión'.

Todas las partes interesadas consultadas reconocieron que el racismo, la xenofobia o la transfobia pueden amplificarse a través de las redes sociales. También reconocieron cómo la moderación de contenidos afecta a la libertad de expresión y los problemas asociados a delegar la moderación de contenidos en empresas privadas.

Un entendimiento más profundo de las prácticas de moderación de contenidos puede aportar nuevas perspectivas sobre problemáticas sociales y propiciar reflexiones sobre dónde deben situarse los límites de la libertad de expresión. Una representante de la organización feminista Artemisas lo explicó así: 'Asuntos como el racismo requieren un

ejercicio pedagógico, sobre todo cuando se borra un tuit, la gente debería saber que se ha borrado un tuit porque era racista u homofóbico, por ejemplo'. Con un enfoque diferente, el Colectivo de Mujeres Wiwas, un grupo de defensa de los derechos de las personas afrocolombianas, mencionó que su estrategia no consiste en denunciar los comentarios racistas, sino en ignorarlos: 'No respondemos a los comentarios de odio, no los denunciamos. Es nuestra forma de evidenciar el racismo'.

Si bien hay consenso en que la moderación de contenidos puede afectar a la libertad de expresión, actualmente no existe un acuerdo entre las distintas partes interesadas sobre los tipos específicos de expresión que se deberían moderar en las redes sociales. Tampoco hay consenso sobre las mejores medidas para abordar esos contenidos más allá de la moderación de contenidos, sin poner en peligro los derechos a la libertad de expresión, la privacidad y la participación política. Incluso entre las distintas organizaciones que conforman la sociedad civil y los diversos organismos que vigilan los compromisos internacionales en materia de derechos humanos, existen posturas divergentes sobre los contenidos que deben circular en las redes sociales y las medidas que deben adoptar los Estados y las plataformas.

Las organizaciones entrevistadas para esta investigación son, por tanto, variadas y reflejan un amplio espectro de perspectivas sobre la moderación y la curación de contenidos. Sus intereses son diversos y sus posturas sobre el cambio son diferentes; algunas de ellas pueden ser incluso contradictorias. Sin embargo, hay consenso en que primero es necesario comprender detalladamente y debatir cómo se produce la moderación y curación de contenidos, cómo funciona la automatización en la moderación, cuáles son las consecuencias para los usuarios y para la libertad de expresión, y quién forma parte del proceso.

En el contexto colombiano, sería beneficioso que una coalición se enfocara inicialmente en entender el papel de las plataformas y el Estado en el debate público y los riesgos asociados a estas. El funcionamiento de los procesos de moderación de contenidos, el impacto y las consecuencias de la práctica desarrollada por las empresas, así como el

papel desempeñado por el gobierno colombiano también se encuentran entre los temas a trabajar por una coalición.

Comprensión compartida del papel del Estado y las plataformas de redes sociales

Una coalición o red sobre moderación de contenidos y libertad de expresión en Colombia no solo debería fomentar las relaciones entre sus miembros, sino desarrollar un diálogo con dos partes interesadas: las plataformas de redes sociales y el Estado.

Las partes interesadas que entrevistamos se relacionan a distintos niveles con las plataformas de redes sociales. El representante de la FLIP indicó que los canales de comunicación con algunas plataformas como Facebook (a través del programa de socios de confianza (*Trusted Partners*)) o X parecen depender de la voluntad política y del interés del personal en ciertos temas. La FLIP ha notado que la comunicación con las plataformas era más eficiente hace unos años, mientras que ahora es más difícil obtener respuesta de las plataformas sobre parte del contenido reportado. La sociedad civil tiene interés en establecer una conversación verdadera con las plataformas y expandir la discusión de ciertos contenidos en concreto a las políticas y las oportunidades para los cambios estructurales.

Para Observacom, la relación entre las plataformas, los usuarios y los investigadores es asimétrica: 'Existe gran cantidad de información relevante para los usuarios e investigadores que se esconde del escrutinio público'. Esto sugiere que las plataformas toman decisiones unilaterales sin entender el contexto y muchas veces sin notificar o responder las apelaciones.

Si los académicos pueden mediar en los debates entre usuarios, en tanto son percibidos como actores independientes en el campo, también pueden desarrollar investigación sobre los impactos de las prácticas de moderación y curación de contenidos si acceden a datos de las plataformas.[38] Sin embargo, su capacidad para hacerlo se ha visto obstaculizada por las plataformas que limitan el acceso a las API (las herramientas que

proporcionan acceso a los datos de la plataforma y, por ende, a lo que pasa en aquel espacio público digital). Las plataformas no solo han limitado los tipos de investigadores que pueden solicitar acceso a las API, sino que la tendencia reciente es incluir *paywall*, lo que restringe mucho más el acceso.

Es claro que se requiere más interacción con las plataformas. Lo anterior, facilitaría la incidencia para mejorar las prácticas de moderación y curación de contenidos y que estas se ajusten al contexto colombiano. Asimismo, propiciaría la transparencia sobre cómo funcionan estas prácticas y de qué manera influencian el debate público en línea en Colombia.

El Estado puede influir significativamente en la manera en que se modera o cura el contenido en redes sociales. Las funciones de regulación en Colombia se encuentran dispersas en distintas autoridades que no tienen claridad interna sobre su relación con las plataformas. MinTIC conoce la relación entre la regulación del contenido en línea y la libertad de expresión. El problema es que, si bien MinTIC es la entidad gubernamental responsable de los asuntos relacionados con la gobernanza de Internet, cuando entrevistamos a sus funcionarios no ofrecieron información exhaustiva sobre la implementación de poderes legales y cómo impactan en el ecosistema digital. Esto es de particular importancia si las regulaciones locales pretenden impactar a las plataformas globales.

Esta investigación evidenció que las entidades colombianas solicitan restricciones de contenido y de cuentas a través de mecanismos legales y de las reglas de la comunidad. Es razonable concluir que los servidores públicos no miden el impacto en derechos humanos de sus decisiones. Estas no son públicas y tienden a no ser revisadas por otras autoridades.

Idealmente, las partes interesadas deben desarrollar estrategias para interactuar con el Estado y así, incidir a favor de decisiones basadas en derechos, fomentando la participación de partes interesadas que se encuentran lejos del diálogo con los tomadores de decisiones. Una coalición sobre moderación de contenidos en Colombia debe

equilibrar los distintos niveles de conocimiento, participación, relevancia e interacción con actores clave. La coalición debería también promover espacios para el intercambio de información.

Necesidades, brechas y fortalezas para una posible coalición

La sociedad civil latinoamericana ha desarrollado múltiples coaliciones y alianzas, así como distintos proyectos conjuntos y producciones que desarrollan o se relacionan con el amplio tema de moderación de contenidos y regulación de las plataformas. Además del trabajo desarrollado por IFEX-ALC, Al Sur, Voces del Sur, Alianza Regional por la Libre Expresión y Acceso a la Información, La Alianza por el Cifrado en América Latina y el Caribe (AC-LAC) y la Coalición Derechos en la Red existe bastante cooperación dentro de los países y en la región con miras a lograr algún tipo de regulación de las plataformas de redes sociales.

Esta investigación evidencia que las coaliciones que existen en asuntos relevantes de moderación de contenidos (derechos digitales, terrorismo, desinformación, rendición de cuentas o violencia de género) son más adecuadas que crear una nueva coalición. Las coaliciones existentes ya han desarrollado un entendimiento común sobre la moderación de contenidos y han realizado peticiones específicas a las plataformas y las entidades del Estado.

Las coaliciones de múltiples partes interesadas con objetivos de incidencia delimitados son un buen modelo para Colombia. Actualmente, distintos actores sociales tienen gran interés en la moderación de contenidos y el modelo de coaliciones existente se percibe como más exitoso en el contexto local.

Si bien la idea de una coalición que se centre exclusivamente en moderación de contenidos y libertad de expresión es elogiable, las investigadoras recomiendan inicialmente hacer un mapeo e interactuar con las coaliciones existentes sobre moderación de contenidos con el fin de desarrollar un modelo alternativo de una ‘red ampliada’. Para ello sería necesario:

1. Identificar temas concretos de incidencia que conecten el trabajo de cada organización con la moderación de contenido. Esto puede hacerse al centrarse en la discusión de una ley, decreto, decisión judicial o campaña en específico. La ley en contra de la violencia basada en género en política es un ejemplo (véase el estudio de caso 3).
2. Fomentar el conocimiento y la conciencia entre los posibles miembros de la red ampliada sobre moderación y curación de contenidos y determinar el impacto del modelo de negocio de las plataformas de redes sociales. Se debe fomentar la investigación y la construcción de capacidades para entender el complejo panorama de las redes sociales y para aclarar cómo las prácticas de moderación o curación de contenido, más allá de la eliminación de contenido o el bloqueo de cuentas, afecta la libertad de expresión en línea y el debate público en Colombia.
3. Desarrollar peticiones comunes que resuenen con los grupos existentes de cara a las acciones de incidencia frente al Estado y las plataformas.

Con el fin de ampliar sus miembros a una variedad de voces, las coaliciones pueden establecer acuerdos (como acuerdos de asociación) para colaborar con otras organizaciones o redes que trabajen en temas que deberían incluirse en el debate de la moderación de contenido, pero que por el momento no trabajan en asuntos de moderación de contenido.

Este formato garantiza la sostenibilidad, dado que las coaliciones existentes ya han desarrollado confianza para trabajar en conjunto. Además, se encuentra estructurado con planes estratégicos y compromisos de dedicación de tiempo a la coalición.

Luego de hacer un mapeo e interactuar con las coaliciones, las investigadoras recomiendan diseñar un acuerdo para esta red ampliada, así como desarrollar pasos concretos hacia el trabajo conjunto sobre bases comunes.

La estrategia propuesta tiene dos aspectos. Primero, las coaliciones existentes se encuentran en una mejor posición para continuar trabajando en moderación de contenidos en tanto ya tienen acuerdos y procesos de toma de decisión. También han

alcanzado cierto nivel de entendimiento mutuo y confianza entre sus miembros y una madurez que les permite articular las necesidades y brechas para expandir el conocimiento y alcance de su incidencia. Segundo, tienen suficiente influencia para crear una red ampliada que incluya organizaciones que actualmente no trabajan en moderación de contenido, como aquellas que trabajan derechos de las comunidades afrodescendientes o indígenas, o de otros grupos vulnerables (como los derechos LGBTIQI+ o derechos de la niñez). Esto podría mitigar el riesgo de que el trabajo se centre demasiado en un tema específico.

Estudio de caso 3: El Observatorio de Violencia Contra las Mujeres en Política

La sociedad civil está conformada por organizaciones con distintos intereses, agendas y experiencias, que tienen sus propias iniciativas de incidencia y prioridades. Si bien la diversidad de puntos de vista es un sinónimo de la riqueza de perspectivas, también significa que frente a ciertos temas existen posiciones contradictorias al interior de una coalición. A medida que crece la coalición, también crecen las posiciones contradictorias. Esta situación debe ser tenida en cuenta al desarrollar una iniciativa que busca atajar los actuales problemas de la moderación y curación de contenidos en Colombia.

La diversidad al interior de las coaliciones no es necesariamente un impedimento, sino que es precisamente la fortaleza de las iniciativas de múltiples partes interesadas. De hecho, varias entidades se han organizado en coaliciones para incidir en ciertas causas con distintos niveles de éxito. La clave para la sostenibilidad se encuentra en compartir un entendimiento común y objetivos similares, construir confianza, desarrollar una estructura y garantizar la disponibilidad de recursos.

La experiencia de una coalición que promueve la expedición de una ley para prevenir y sancionar la violencia contra las mujeres en política (que incluye disposiciones sobre moderación de contenidos) es un buen estudio de caso. El Observatorio de Violencia Contra las Mujeres en Política es una red de actores nacionales e internacionales que trabajan en el monitoreo y análisis de la violencia contra las mujeres en política en Colombia.[39] Hacen parte de esta alianza entidades estatales y organizaciones que trabajan en asuntos de violencia política, elecciones, derechos de las mujeres, derechos digitales, etc.

El Observatorio fue el artífice de la reciente campaña de incidencia que llevó a la expedición de una ley en 2023 que incluye disposiciones para contrarrestar la violencia política en línea. El texto legislativo regula el asunto de manera comprensiva e incluye sanciones y deberes de prevención y capacitación a cargo de entidades públicas y privadas. Luego de múltiples intentos, se trata de la primera regulación exitosa de la moderación de contenidos en Colombia.

El Observatorio en primer lugar, llegó a un acuerdo sobre cómo se entiende el asunto principal que pretendía enfrentar: la violencia política contra las mujeres. Sin embargo, algunos miembros del Observatorio tenían puntos de vista distintos sobre cómo enfrentar la violencia en línea. Algunos tenían preocupaciones basadas en la libertad de expresión y se oponían a las disposiciones que le daban poderes legales a las autoridades públicas para solicitar la remoción del contenido que aparece en redes sociales sin límites claros relativos al tipo de contenido que puede eliminarse y la legitimidad para hacerlo. Por otro lado, las organizaciones que trabajaban de la mano con mujeres víctimas de violencia hacían un llamado urgente por restricciones más estrictas en el tipo de contenido que se permite en redes sociales debido a los efectos que tiene en la vida real. Aunque el encuentro entre estas dos posiciones produjo discusiones desafiantes e interesantes, e incluso a veces reveló que existían fricciones entre aliados, centrarse en el debate de un asunto en particular ayudó a los miembros del Observatorio a llegar a compromisos mutuos y a soluciones intermedias.

Una red ampliada que trabaje en asuntos no directamente relacionados con la moderación de contenidos

Algunas partes interesadas de gran importancia en el debate sobre moderación de contenidos todavía no trabajan en asuntos de moderación y curación de contenido, como las organizaciones que trabajan en los derechos de las comunidades afrodescendientes o indígenas, o de otros grupos vulnerables (como los derechos LGBTIQI+ o derechos de la niñez). Sin embargo, existen varias coaliciones que se centran en estos temas en Colombia. Como evidencia la investigación, estas comunidades son objeto de violencia estructural en entornos digitales, especialmente en redes sociales, y son una parte esencial de la discusión sobre moderación de contenido.

Las representantes de las organizaciones que trabajan en estos asuntos entrevistadas se mostraron interesadas en la moderación de contenido, a pesar de que su nivel de comprensión fuera sustancialmente menor comparado al de las organizaciones que ya trabajan en derechos digitales y regulación de plataformas. También compartieron preocupaciones similares sobre la necesidad de más transparencia de las plataformas y sobre las prácticas de moderación de contenido.

Sin embargo, no estaba claro cuán probable era que experimentaran problemas de moderación de contenido, lo que conlleva el riesgo de que estas organizaciones no sigan, dediquen recursos o contribuyan a la red ampliada sobre moderación de contenido. Es esencial incluir estas voces y fortalecer su incidencia a través del desarrollo de capacidades en asuntos de libertad de expresión y cómo esto aplica al espacio en línea, especialmente sobre prácticas de moderación de contenido, el funcionamiento del sistema desarrollado por las empresas de redes sociales y el impacto de su modelo de negocios. Incrementar las habilidades para comprender la moderación de contenidos al interior de estas organizaciones les permitirá identificar su interés en la agenda de la regulación de la moderación de contenido. También ayudará a que establezcan intereses en común con quienes ya trabajan en agendas similares.

El mejor camino hacia el éxito es establecer una amplia red que pueda desarrollarse en distintas fases y crecer a medida que su número de miembros aumenta. Una red ampliada permitirá la colaboración de organizaciones o coaliciones al desarrollar acuerdos de

colaboración con objetivos, responsabilidades y actividades claras, como desarrollo de capacidades, investigación, incidencia, realizar campañas o sensibilizar al público dependiendo de los intereses comunes.

Análisis de las partes interesadas

Esta sección presenta las coaliciones de múltiples partes interesadas existentes que pueden participar en el tema de moderación de contenidos en Colombia,

El Índice de Derechos Digitales

Se trata de una alianza de múltiples partes interesadas compuesta por organizaciones de diversos contextos (academia, derecho y tecnología, periodismo y análisis de datos): El Veinte, FLIP, Fundación Karisma, ISUR, Linterna Verde, y Dejusticia. Uno de los temas estudiados por la alianza es el control de contenidos, incluyendo el análisis de la moderación de contenidos que hacen las plataformas y las peticiones gubernamentales durante la pandemia de Covid-19.

Ventajas:

- tiene un enfoque multiparte;
- se enfoca en los derechos digitales y su misión se relaciona con la libertad de expresión;
- mantiene canales de comunicación abiertos con las empresas de redes sociales y con el Estado en distintos asuntos;
- tiene pocos miembros, lo que facilita la toma de decisiones para iniciar procesos; y
- su misión puede dar cabida a la moderación de contenido.

Posibles temas de interés:

- moderación de contenidos y libertad de expresión; y
- transparencia algorítmica, transparencia de las plataformas e intercambio de información.

Red de Acción Cívica contra la Desinformación (ACD)

Esta red, gestionada por CIVIX, se compone de organizaciones colombianas interesadas en la 'desinformación' e incluye a entidades públicas. La 'desinformación' es un tema que evidentemente se relaciona con la moderación de contenido. ADC incluye a medios de comunicación como La Silla Vacía, ColombiaCheck, El Mundo, Prensa Escuela y El Universal, organizaciones no gubernamentales como Hablemos, Dividendo por Colombia y Fundación Carvajal y representantes del sector académico como la Universidad Nacional de Colombia y varias Secretarías de Educación municipales.

Ventajas:

- tiene un enfoque multiparte;
- interesada en la 'desinformación' que se relaciona con la moderación de contenidos y la libertad de expresión; y
- cuenta con un gran número de miembros y de distintas partes interesadas, lo que permite múltiples perspectivas y posiciones ampliamente respaldadas en las que se alcanza el consenso.

Posibles temas de interés:

- el rol de las plataformas en el ecosistema de la información; y
- transparencia algorítmica, transparencia de las plataformas e intercambio de información.

Observatorio de Violencia contra las Mujeres en Política

El Observatorio es una alianza muy diversa integrada por organizaciones de la sociedad civil que trabajan en los ámbitos de los derechos digitales, los derechos electorales y los derechos de la mujer, junto con autoridades estatales y organizaciones internacionales. El Observatorio está compuesto por el Ministerio de Justicia de Colombia, la Consejería Presidencial para la Equidad de la Mujer, ONU Mujeres, Transparencia por Colombia, la Secretaría de la Mujer de la Alcaldía de Bogotá, el Instituto Holandés para la Democracia

Multipartidaria (NIMD), el Instituto Nacional Demócrata, el Consejo Nacional Electoral y la Fundación Karisma.

El proyecto de ley sobre violencia contra las mujeres en la vida política fue aprobado por el Congreso y actualmente se encuentra ante la Corte Constitucional para su revisión automática (se espera que la Corte lo revise en el primer semestre de 2024). El trabajo del Observatorio ofrece una oportunidad clara para desarrollar una regulación sobre moderación de contenidos en Colombia.

Ventajas:

- es la coalición más amplia en Colombia e incluye a diversas partes interesadas, como a autoridades estatales y al sector privado;
- logró que el Congreso aprobara la primera regulación en el país relacionada con la moderación de contenidos en las plataformas de redes sociales; y
- trabajará en el tema de moderación de contenidos en los próximos meses.

Posibles temas de interés:

- desafíos en la implementación, buenas prácticas y otras preocupaciones (casos de estudio, especialmente México); y
- transparencia algorítmica, transparencia de las plataformas e intercambio de información.

Alianza por la igualdad de las mujeres en los medios

Esta alianza incluye periodistas, académicas, medios de comunicación y otras organizaciones de la sociedad civil interesadas en la igualdad de las mujeres en los medios. También tiene fuertes lazos con entidades interesadas en la violencia de género. La alianza está conformada por la Red de periodistas con visión de género, FLIP, Sentido, Colnodo, Fundación Karisma, Consejo de Redacción y Línea del Medio.

Debido a que la Corte Constitucional exhortó al Congreso a llenar un vacío legal, los debates sobre el proyecto de ley en materia de igualdad de género en los medios serán parte de la agenda legislativa de los próximos meses. Este proyecto de ley tendrá en cuenta las disposiciones del proyecto de ley sobre la violencia contra las mujeres en la vida política y probablemente incluirá la moderación de contenidos en las plataformas de redes sociales.

Apoyar a esta alianza ofrece otra clara oportunidad para desarrollar una legislación sobre moderación de contenidos basada en derechos, si las partes interesadas lo consideran pertinente.

Ventajas:

- cuenta con información sobre violencia de género en redes sociales y rápidamente puede ofrecer evidencia y apoyo a la redacción del proyecto de ley sobre mujeres periodistas;
- el exhorto de la Corte al Congreso le da fuerza a su labor de incidencia;
- en los próximos meses estará trabajando en la regulación sobre moderación de contenidos en virtud de su interés por la violencia digital contra mujeres periodistas; y
- tiene pocos miembros.

Posibles temas de interés:

- desafíos en la implementación, buenas prácticas y otras preocupaciones; y
- transparencia algorítmica, transparencia de las plataformas e intercambio de información.

Conclusión

Durante el transcurso de la investigación, se identificaron las partes interesadas y los temas relevantes en la intersección entre la moderación de contenidos en línea y la libertad de expresión en Colombia. Se entrevistó a las partes interesadas para obtener una imagen realista y completa de cómo las redes sociales moderan y curan el contenido en Colombia y el impacto que esto tiene sobre la libertad de expresión en el país. El informe también analizó la viabilidad de establecer una coalición local sobre moderación de contenidos y libertad de expresión. Concluye que la mejor estrategia consiste en involucrar a las coaliciones existentes que trabajan en temas relacionados con la moderación de contenidos y apoyarlas para que amplíen sus objetivos y abarquen sus desafíos. El objetivo de la coalición podría ser garantizar que la moderación de contenidos en Colombia se ajuste a las normas internacionales sobre libertad de expresión y al contexto local.

La experiencia de las organizaciones de la sociedad civil que han participado en este estudio es muy amplia, sus intereses diversos y sus apuestas por el cambio diferentes; algunas de ellas pueden ser incluso contradictorias. El resultado es un amplio espectro de perspectivas sobre la moderación y la curación de contenidos. Sin embargo, todas las partes interesadas entrevistadas subrayaron la importancia de llegar a un entendimiento común sobre el funcionamiento de la moderación de contenidos, el funcionamiento de la automatización, las consecuencias para los usuarios y para la libertad de expresión, y la responsabilidad de estos procesos.

Actualmente, la sociedad civil, la academia y los medios de comunicación no coinciden en una serie de cuestiones. Entre ellas, un enfoque común de los contenidos 'perjudiciales' amparados por la libertad de expresión, incluso entre ellos mismos; las medidas para abordarlos de la mejor manera, sin poner en peligro los derechos a la libertad de expresión, la privacidad y la participación política; y las medidas que deberían adoptar los Estados y las plataformas en materia de moderación de contenidos. La comprensión de lo que puede constituir un contenido 'perjudicial' -inherente a la naturaleza abierta del

significado del término 'perjudicial'- también puede estar sujeta a cambios debido a las diferencias sociales entre las partes interesadas y los usuarios de las plataformas en términos de religión, creencias políticas y contexto específico.

Con todo, existe una fuerte posición común de que la sociedad civil colombiana tiene una cultura de libertad de expresión profundamente arraigada, y todas las partes interesadas entrevistadas reconocieron que la moderación de contenidos puede suponer un riesgo para la libertad de expresión que debe ser atendido.

Se observó que la labor de la sociedad civil se ve a menudo obstaculizada por prácticas de moderación o curación de contenidos, por ejemplo, al denunciar abusos de las autoridades estatales. En estos casos, las partes interesadas culpan a las plataformas o al Estado y consideran esas prácticas como herramientas de censura. Las partes interesadas no suelen comprender las diferentes políticas de contenidos y los procesos de aplicación de estas. Los estudios de caso y los testimonios muestran cómo la moderación de contenidos puede llevar a silenciar voces, lo que puede dar lugar a un 'agujero negro digital'.

Todas las partes interesadas entrevistadas compartían la misma queja sobre la opacidad de las normas comunitarias de las plataformas y su proceso de moderación de contenidos y toma de decisiones. Las partes interesadas exigen más transparencia tanto en la toma de decisiones como en los procesos, y que las plataformas proporcionen más información sobre las prácticas de moderación de contenidos a través del contexto local.

Otra preocupación emergente son las prácticas de curación de contenidos. Éstas se perciben como especialmente difusas y alarmantes, y los entrevistados perciben sus efectos como censura. Es necesario seguir investigando y desarrollando capacidades para comprender si las nuevas formas de curación de contenidos producen este efecto o sensación, o si se confunden con las prácticas de moderación de contenidos.

Teniendo en cuenta que las redes sociales son actores clave en el ecosistema de la información, su impacto e influencia, este informe ha demostrado que todavía hay una

falta de conocimiento contextual entre las partes interesadas sobre el funcionamiento y el impacto de la moderación y la curación de contenidos en Colombia. Incluso si las plataformas tratan de proporcionar información y explicaciones, no es suficiente para explicar eficazmente si tienen en cuenta las especificidades locales y si sus procesos pueden afectar a la realidad del discurso en línea en Colombia. Por ejemplo, aunque las empresas de redes sociales indican que cumplen las leyes de los países en los que operan, no está claro hasta qué punto las leyes nacionales se aplican o afectan a sus decisiones; este vacío impide que las partes interesadas tengan posiciones más informadas sobre la regulación.

De la respuesta escrita recibida por el MinTIC se desprende que el Ministerio es consciente de la relación entre la regulación de los contenidos en línea y la libertad de expresión. Aunque el MinTIC es la entidad gubernamental responsable de los asuntos de gobernanza de Internet, su declaración escrita muestra que la falta de análisis interno sobre la regulación actual es una oportunidad perdida para recopilar lecciones aprendidas e información sobre el papel del Estado en caso de que se amplíen sus facultades legales a través de regulaciones locales que apliquen a las plataformas. Es posible concluir que los funcionarios no miden el impacto de estas decisiones en los derechos humanos, y que existe un gran vacío, ya que las decisiones no son públicas ni están sujetas a revisión posterior por parte de otra autoridad.

La investigación también ha demostrado que la moderación y la curación de contenidos cumplen una función importante en el proceso de consolidación de la paz en Colombia. Por ejemplo, el contenido que puede contener material violento podría infringir las normas de la comunidad en diferentes plataformas, pero al mismo tiempo puede ser importante para el debate público por su capacidad de constituirse como prueba de abusos estatales o por su papel en la construcción de memoria. Se han hecho numerosos llamados para mantener este tipo de contenidos visibles en el espacio digital público.

Con el fin de establecer un mecanismo que se ocupe de estos asuntos, esta investigación concluye que una coalición de partes interesadas más amplia compuesta por

organizaciones de distinta procedencia y competencias desiguales no funcionaría en el contexto colombiano. En cambio, se propone construir sobre las coaliciones existentes que ya trabajan en asuntos relacionados con la moderación de contenidos, expandiendo sus objetivos y conectándolas con los procesos regulatorios actuales. Esta estructura se puede definir como una 'red ampliada', liderada por las coaliciones existentes con una estructura y gobernanza establecidas, junto a organizaciones o coaliciones que trabajan en asuntos que no se relacionan directamente con la moderación de contenidos, pero que representan partes interesadas importantes (por ejemplo, organizaciones que trabajan en los derechos de las minorías o de las comunidades indígenas, las mujeres, la población LGBTQI+, etc.).

Recomendaciones

Las siguientes recomendaciones pretenden orientar sobre los próximos pasos hacia la conformación de una red que se enfoque en la moderación de contenidos y la libertad de expresión en Colombia.

Debates sobre moderación de contenidos

Las discusiones sobre moderación de contenidos entre algunas partes interesadas son complejas. Estas pueden alinearse con oportunidades regulatorias donde las partes interesadas se sientan fuertemente impactadas por la moderación de contenidos. Comprender los puntos ciegos que afectan la libertad de expresión en línea (sobre todo aquellos relacionados con la moderación de contenidos) como parte de las discusiones regulatorias más amplias puede ser un impulso para involucrar a las coaliciones existentes en el trabajo previsto por el proyecto Social Media 4 Peace en Colombia.

Para aprovechar las futuras discusiones regulatorias en Colombia, la 'red ampliada' puede enfocar su trabajo y estrategias en:

1. Abogar por una mayor transparencia de las plataformas de redes sociales sobre las prácticas de moderación de contenidos;
2. Realizar investigación sobre el impacto de las nuevas formas de moderación y curación de contenidos en la libertad de expresión;
3. Realizar investigación en el fenómeno de los 'agujeros negros digitales'.

Objetivo común

La experiencia de las coaliciones exitosas en Colombia muestra que, a pesar de la diversidad de sus posiciones (o tal vez gracias a ella), su fortaleza se debe a que se enfocan en un objetivo de incidencia común o en un tema en específico, como ocurre en procesos legislativos determinados.

Para ser exitosa, la 'red ampliada' debe diseñar un objetivo común que recoja las posiciones de sus miembros. Esto podría incluir solicitudes específicas a ciertos actores clave, como las plataformas o el Estado, sobre asuntos de moderación de contenidos en momentos clave.

Desarrollo de capacidades y conocimiento

Las partes interesadas en Colombia están de acuerdo en que las plataformas de redes sociales no ofrecen suficiente información sobre sus procesos de moderación y curación de contenidos. También hay críticas relacionadas con la vaguedad de las normas comunitarias. Construir sobre estas quejas y preocupaciones compartidas es clave para garantizar la participación de las organizaciones o coaliciones que actualmente no centran su labor directamente en la moderación de contenidos.

Con este fin, debe reforzarse la capacidad y los conocimientos de las posibles partes interesadas de la propuesta 'red ampliada' en las prácticas de moderación de contenidos y su impacto en la libertad de expresión. La capacitación y el intercambio de conocimiento también deberían centrarse en las obligaciones de las plataformas de redes sociales y en una mejor comprensión de los problemas que tienen los informes de transparencia que se

producen hoy en día. Esto dotaría a las partes interesadas con más argumentos para exigir la transparencia sobre la moderación de contenidos y para pedir una participación más activa del Estado.

Colaboración

La 'red ampliada' debería enfocarse en crear caminos que permitan la colaboración con las plataformas de redes sociales para generar un diálogo que contribuya a resolver las fallas de la moderación y curación de contenidos y la protección de los derechos fundamentales. La interacción con las plataformas debería ir más allá de la resolución de casos concretos para abordar las estructuras y procesos de las plataformas, el impacto de su modelo de negocio y su funcionamiento en Colombia.

La población general tiene pocos conocimientos sobre la importancia y funcionamiento de la moderación y curación de contenidos. Si bien el Estado tiene la responsabilidad de promocionar las habilidades y los conocimientos de los ciudadanos en términos de una mayor alfabetización digital e informacional, las plataformas deben garantizar que sus usuarios comprendan claramente sus normas comunitarias. La 'red ampliada' de partes interesadas podría tener un rol activo en el trabajo para lograr estos objetivos y podría incidir a favor de un programa de alfabetización digital desarrollado por el Estado o las plataformas, o incluso participar en él.

Investigación

La sociedad civil y la academia necesitan acceder a los datos de las plataformas a través de APIs gratuitas para producir investigaciones que vayan más allá de los estudios de caso e incluya aspectos clave de la moderación de contenidos. La futura 'red ampliada' podría adoptar este objetivo en su labor de promoción ante las plataformas de redes sociales.

Anexo A: Análisis de riesgos

La **Coalición sobre Libertad de Expresión en Línea y Moderación de Contenidos** surge como una oportunidad única para la participación y contribución de todos los actores y como un mecanismo de profundo cambio. La coalición ofrece un camino hacia el consenso en asuntos clave de moderación de contenidos y las oportunidades para enfrentarlos. La siguiente tabla ofrece una visión general de los posibles riesgos identificados por las personas entrevistadas relacionados con la formación y funcionamiento de la coalición, incluyendo los caminos para superarlos y mitigarlos.

Tipo de riesgo*	Descripción del riesgo	Probabilidad**	Impacto***	Monitoreo y mitigación
Institucional	Tiempo y recursos humanos de las organizaciones para dedicar al trabajo de la coalición sobre moderación de contenidos.	Probable	Menor	• Aunque las organizaciones disponen de escasos recursos para dedicarlos a nuevos empeños o iniciativas, la moderación de contenidos forma parte de la agenda de algunas. • Para tener éxito, cualquier iniciativa tiene que capitalizar el trabajo organizativo existente y aprovecharlo.
Institucional	Llegar a acuerdos entre los miembros de la coalición y generar confianza.	Probable	Mayor	• Si se aprovechan las coaliciones existentes, los acuerdos, los procesos de toma de decisiones y la confianza ya están establecidos. La integración de cualquier nuevo miembro debe seguir un proceso que garantice que los acuerdos de la coalición sean conocidos y aceptados por todos los miembros. • Es necesario establecer acuerdos sobre la incidencia y posturas sobre moderación de contenidos. Deben

				preverse actividades para el desarrollo de capacidades y la consolidación de la red mientras se establece una red ampliada.
Financiero	La sostenibilidad y longevidad de la coalición dependen de la disponibilidad de fondos.	Probable	Mayor	• Financiación sostenible para la coordinación de la coalición y la consolidación de sus actividades. • Solicitud conjunta de financiación y participación de los miembros de la coalición en las reuniones de donantes y en el proceso de elaboración de las agendas.
Político	Participación efectiva en las decisiones de política pública sobre moderación de contenidos.	Poco probable	Menor	• La coalición puede efectivamente asistir y ser consultada para la formulación de política pública, incidiendo por los valores y objetivos que deba garantizar una potencial regulación futura.
Institucional	Acuerdos entre organizaciones para la participación en la coalición.	Probable	Mayor	• La naturaleza de las organizaciones es diversa; pueden existir perspectivas opuestas frente a ciertos temas. Las reuniones deben contar con metodologías que enlacen opiniones diferentes en la conversación.

Notas:

* El tipo de riesgo se clasifica según las siguientes categorías: Institucional, Político, De Seguridad, De Partes Interesadas, Financiero, *Compliance*, Reputacional, Otro y Covid-19.

** La probabilidad del riesgo se presenta en una escala: Poco probable, Posible, Probable y Casi cierto.

*** El impacto del riesgo se presenta en una escala: Menor, Moderado, Mayor y Severo.

Anexo B: Ficha de entrevistas

Las investigadoras realizaron entrevistas con representantes de las siguientes organizaciones:

Organización	Nombre de la persona entrevistada	Categoría	Tema
–	Luisa Isaza	Investigadora de la Universidad de Oxford	Especialista en libertad de expresión en Internet
Artemisas	Juliana Herrera	Sociedad civil	Derechos de las mujeres
Caracol TV	Anónimo	Medio de comunicación	Canal de televisión colombiano
Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE)	Agustina Del Campo	Academia o centro de estudios	Investigación sobre libertad de expresión y acceso a la información
Centro Plurales UR	Flora Rodriguez	Academia o centro de estudios	Centro para la Diversidad, Equidad e Inclusión de la Universidad del Rosario
Colectivo Wiwas	Cindy Pérez	Colectivo afro	Difusión de los conocimientos afroculturales de la comunidad Wiwa
ColombiaCheck	Ana Saavedra	Medio de comunicación	Fact-checking de contenidos en línea
Échele Cabeza	Julián Quintero	Sociedad civil	Reducción de riesgos y daños por el consumo de sustancias psicoactivas

El Veinte	Ana Bejarano	Sociedad civil	Libertad de expresión y litigio estratégico
Externado University	Anónimo	Academia	Solicitud de anonimato
Facebook	Grupal	Plataforma digital	Plataforma
Federación Colombiana de Periodistas (FECOLPER)	Jorge Velásquez	Sociedad civil	Red colombiana de periodistas
FLIP	Jonathan Bock	Sociedad civil	Libertad de expresión y libertad de prensa
Fundación Gabo	Ricardo Corredor	Sociedad civil	Periodismo
Fundación Interpreta	–	Sociedad civil	Investigación sobre problemas sociales complejos
Fundación Santamaria	Kika Ruiz	Sociedad civil	Derechos trans
Google	Grupal	Plataforma digital	Plataforma
Indepaz	Juana Cabezas	Sociedad civil	Construcción de paz en Colombia
La Liga Contra el Silencio (League Against Silence)	Alejandro Gómez	Red de medios de comunicación	Historias censuradas en Colombia
La Silla Vacía	Daniel Pacheco	Medio de comunicación	Noticias, historias y debates sobre el poder en Colombia
Linterna Verde	Carlos Cortes	Sociedad civil	Opinión pública en espacios digitales
Ministerio de las Tecnologías de la	Documento escrito firmado por Aylin	Institución pública	Responsable de las tecnologías de la información

Información y las Comunicaciones	Torregroza Villarreal (Viceministerio de Transformación Digital)		y la comunicación en Colombia
Observacom	Gustavo Gómez	Centro de estudios	Regulación y políticas públicas relacionadas con los medios de comunicación, las telecomunicaciones, Internet y la libertad de expresión
ONIC	Wilson Herrera	Red indígena	Organización Nacional Indígena
Red Papaz	Carolina Piñeros	Organización sin ánimo de lucro	Protección de la infancia en el entorno digital
Sentiido	Lina Cuellar	Medio digital	Género, diversidad y cambio social
Temblores ONG	Alejandro Lanz	Sociedad civil	Transformación social
Wikimedia Colombia	Mónica Bonilla	Sociedad civil	Educación y acceso al conocimiento mediante herramientas digitales
X	Grupal	Plataforma digital	Plataforma

Anexo C: Políticas de contenidos de las plataformas principales

Meta	X	YouTube	TikTok
Conducta delictiva y violencia • Violencia e incitación • Personas y organizaciones peligrosas • Organización de actos dañinos y fomento de actividades delictivas • Bienes y servicios restringidos • Fraude y engaño	**Seguridad** • Discurso violento • Entidades violentas y de incitación al odio • Explotación sexual infantil • Abuso/acoso • Conducta de odio • Perpetradores de ataques violentos. • Suicidio • Contenido multimedia delicado • Productos o servicios ilegales o ciertos productos o servicios regulados	**Spam y prácticas engañosas** • Spam, prácticas engañosas y estafas • Robo de identidad • Vínculos externos • Participación falsa • Listas de reproducción • Políticas adicionales	**Seguridad y civismo** • Comportamientos violentos y actividades delictivas • Discursos y comportamientos de odio • Organizaciones e individuos que fomenten la violencia o el odio • Explotación y abuso de menores • Explotación sexual y violencia de género • Explotación humana • Intimidación y acoso
Seguridad • Suicidio y autolesiones • Explotación sexual, abuso y desnudos de menores • Explotación sexual de adultos • Bullying y acoso • Trata de personas • Infracciones de privacidad	**Privacidad** • Información privada • Desnudez no consensuada • Compromiso de la cuenta	**Contenido sensible** • Imágenes de desnudos y contenido sexual • Miniaturas • Seguridad infantil • Suicidio y autolesiones • Lenguaje vulgar	**Salud mental y conductual** • Suicidio y autolesiones • Trastornos alimentarios e imagen corporal • Retos y actividades peligrosas

Conducta reprochable • Lenguaje que incita al odio • Contenido gráfico y violento • Desnudos y actividad sexual de adultos • Servicios sexuales	**Autenticidad** • Manipulación de plataforma y spam • Integridad cívica • Identidades engañosas • Contenido multimedia falso y alterado	**Contenido violento o peligroso** • Contenido perjudicial o peligroso • Contenido violento o gráfico • Organizaciones criminales violentas • Incitación al odio o a la violencia • Acoso y cyberbullying	**Temas delicados y para adultos** • Actividad y servicios sexuales • Desnudez y exposición del cuerpo • Contenido sexualmente sugerente • Contenido perturbador y gráfico • Abuso animal
Integridad y autenticidad • Integridad de las cuentas y autenticidad de la identidad • Spam • Ciberseguridad • Comportamiento no genuino • Información errónea • Cuentas conmemorativas		**Bienes regulados** • Venta de bienes y servicios ilegales o regulados • Armas	**Integridad y autenticidad** • Desinformación • Integridad de procesos cívicos y electorales • Medios sintéticos y manipulados • Interacciones falsas • Contenido no original y códigos QR • Spam y comportamientos de cuenta engañosos
Respeto de la propiedad intelectual • Propiedad intelectual		**Información errónea** • Políticas de información errónea • Políticas de información errónea durante las elecciones • Políticas de información médica errónea sobre Covid-19	**Bienes regulados y actividades comerciales** • Juegos de azar • Alcohol, tabaco y drogas • Armas de fuego y armas peligrosas • Comercio de bienes y servicios regulados

		• Política de desinformación sobre vacunas	• Comunicaciones comerciales y promoción de pago • Fraudes y estafas

Bibliografía

Access Now (2021) What You Need to Know about the Facebook Papers [Lo que necesitas saber sobre los 'Facebook Papers'].

Adam, M. (2021, 7 de mayo) [@mosseri]. 'Yesterday We Experienced a Technical Bug' [Ayer experimentamos un problema técnico]. Publicación de X.

ARTICLE 19 (2018) Side-stepping Rights: Regulating Speech by Contract, Policy Brief [Eludiendo los derechos: La regulación del discurso por contrato].

ARTICLE 19 (2021) Social Media Councils: One Piece in the Puzzle of Content Moderation [Consejos de redes sociales: una pieza en el rompecabezas de la moderación de contenidos].

ARTICLE 19 (2023) Vigilar a los vigilantes: Moderación de contenido, gobernanza y libertad de expresión.

ARTICLE 19 (2022) Content Moderation and Local Stakeholders in Bosnia and Herzegovina [Moderación de contenidos y partes interesadas locales en Bosnia y Herzegovina].

ARTICLE 19 (n.d.) #MissingVoices.

Asamblea Nacional Constituyente (1991) Constitución Política de Colombia.

Banco Mundial (2020) Población en Colombia.

Botero, C. (2021) Public Memory and the Digital Black Hole [Memoria pública y el agujero negro digital].

Botero, C. (2021) Represión en la calle, sensación de censura en redes.

CELE (2022) Penar la intolerancia 'male sal'. Críticas a la Convención Interamericana contra toda forma de Discriminación e Intolerancia.

Cepeda, M.J. (1995) El derecho a la constitución en Colombia. entre la rebelión pacífica y la esperanza.

Charry, C. (2021) Los en vivo: Estar vivos y ser vistos, Punto y Coma.

Cifras y Conceptos (2023) Sexto estudio de percepción de jóvenes.

Cinep (2008) Comunicación y conflicto armado: El fin no justifica a los medios.

Comisión de la Verdad (2022) Hallazgos y Recomendaciones: Hallazgos y recomendaciones de la Comisión de la Verdad de Colombia.

Comisión de la Verdad (2022) Hay un futuro si hay verdad.

Comisión de la Verdad (2022) No matarás: Relato histórico del conflicto armado interno en Colombia.

Comisión de Regulación de las Comunicaciones (2022) Data Flash 2022-026: Internet Fijo.

Comisión Interamericana de Derechos Humanos (2021) Observaciones y recomendaciones: Visita de trabajo a Colombia.

Comisión Internacional de Juristas (2023) Colombia: Defensores de derechos humanos continuaron bajo presión y ataques.

Comisión para la Eliminación de la Discriminación contra la Mujer (2020) Recomendación general núm. 38, relativa a la trata de mujeres y niñas en el contexto de la migración mundial.

Congreso de la República (2008) Ley 1257 de 2008.

Consejo Asesor de Contenido de Facebook (2021) Manifestaciones en Colombia.

Constitución Política de Colombia (1991) Art. 2, 3 de julio de 1991.

Corte Constitucional (2023) Estadísticas de tutelas radicadas en la Corte Constitucional.

Corte Constitucional de Colombia (2004) Sentencia T 1191 de 2004.

Corte Constitucional de Colombia (2005) Sentencia T-1062 de 2005.

Corte Constitucional de Colombia (2012) Sentencia T-627 de 2012.

Corte Constitucional de Colombia (2015) Sentencia T-066 de 2015.

Corte Constitucional de Colombia (2020) Sentencia T-031 de 2020.

Corte Constitucional de Colombia (2021) Sentencia T-146 de 2021.

Corte Constitucional de Colombia (2022) Sentencia T-280 de 2022.

Corte Constitucional de Colombia (2023) Sentencia T-087 de 2023.

Corte Interamericana de Derechos Humanos (2018) Caso Isaza Uribe v. Colombia.

DANE (2018) Grupos étnicos: Información técnica. Consultado el 1 de diciembre de 2023.

Dejusticia (2021) Homenaje a la tutela: El mecanismo que democratizó la Constitución de 1991.

Del Campo, A. (2022) Contenido legal pero dañino y poca previsión en la supervisión. CELE.

El Economista (2022) El Gobierno anunció un proyecto para regular las redes sociales para 'que dejen de intoxicar' a la democracia.

Fiorella, G. (2019) El segundo a segundo del disparo que mató a Dilan Cruz.

Fitzgerald, M., (2022) No es solo contra Francia: En política, los insultos son contra todas, Revista Cambio.

France24 (2022) Evolucionar: El último giro de la desinformación electoral en Colombia.

Freedom in the World Index (2023) Freedom in the World 2023: Colombia [Libertad en el mundo 2023: Colombia].

Frithjof, S.-M., Britta, H. and Melanie, V. (2012) How Stressful is Online Victimization? Effects of Victim's Personality and Properties of the Incident [¿Qué tan estresante es la victimización en línea? Efectos de la personalidad de la víctima y las propiedades del incidente], *European Journal of Developmental Psychology*, 9, 2: 260–274.

Fundación Karisma (2014) Violencia Contra Las Mujeres Y Tic (Vcm Y Tic).

Fundación Karisma (2021) El proyecto de Ley 600 sigue su curso en el Congreso a pesar de las críticas.

Fundación Karisma (2021) Fallas de internet, bloqueos de redes y censura de contenidos en protestas: Realidades y retos para el ejercicio de los derechos humanos en los contextos digitales.

Fundación Karisma (2021) Periodistas sin acoso.

Fundación Karisma (2021) Pistolas contra celulares.

Fundación Karisma (2021) Violencias machistas atacan la libertad de expresión de periodistas y comunicadoras en Colombia.

Fundación Karisma (2022) Detección automática de derechos de autor: una herramienta de desigualdad.

Fundación Karisma (2022) Dónde están mis datos.

Fundación Karisma, Fundación Para la Libertad de Prensa (FLIP), El Veinte e ISUR (2022) Comentarios al Proyecto de Ley 318.

Fundación Karisma, Fundación Para la Libertad de Prensa (FLIP), El Veinte e ISUR (2023) Comentarios PL (Medidas para prevenir, atender, rechazar y sancionar la violencia contra las mujeres).

Gago, E. (2022) La polarización como estrategia política.

Global Disinformation Index (2022) Disinformation Risk Assessment: The Online News Market in Colombia [Evaluación del riesgo de desinformación: El mercado de noticias en línea en Colombia].

Google (2022) YouTube Community Guidelines Enforcement [Cumplimiento de las normas comunitarias de YouTube].

Google (2023) About the YouTube Trusted Flagger Programme [Acerca del programa de YouTube Trusted Flagger].

Google (2023) Government Requests to Remove Content [Solicitudes gubernamentales de retirada de contenidos].

GSMA Intelligence (2022) Get Started Now with GSMA Intelligence [Empieza ahora con GSMA Intelligence].

Human Rights Watch (2020) 'Video Unavailable': Social Media Platforms Remove Evidence of War Crimes [Video no disponible': Las plataformas de las redes sociales eliminan material sobre crímenes de guerra].

Indepaz (2019) Los discursos del odio y la estigmatización fatal.

Infobae (2021) Ministerio de Defensa confirmó la militarización en Cali: llegan 450 soldados.

Instagram (s.f.) Normas comunitarias.

Instagram (s.f.) Qué hacer si crees que Instagram no debería haber retirado tu publicación.

Comisión Interamericana de Derechos Humanos, Relatoría Especial para la Libertad de Expresión (2010) Marco jurídico interamericano sobre el derecho a la libertad de expresión.

Comisión Interamericana de Derechos Humanos, Relatoría Especial para la Libertad de Expresión (2017) Estándares para una Internet libre, abierta e incluyente.

Comisión Interamericana de Derechos Humanos, Relatoría Especial para la Libertad de Expresión (2021) La desinformación y la libertad de opinión y de expresión.

Comisión Interamericana de Derechos Humanos, Relatoría Especial para la Libertad de Expresión (2022) Mujeres periodistas y salas de redacción.

Informe del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (2017) Promoción, protección y disfrute de los derechos humanos en Internet: medios de cerrar la brecha digital entre los géneros desde una perspectiva de derechos humanos.

Kari, P. (2019) Climate Misinformation on Facebook 'Increasing Substantially', Study Says [La desinformación sobre cambio climático en Facebook 'aumenta sustancialmente', según un estudio], *The Guardian*.

Kepios (2022) We are Social, Digital 2022 Colombia.

Kouzy, R. (2020) Coronavirus Goes Viral: Quantifying the Covid-19 Misinformation Epidemic on Twitter [El coronavirus se vuelve viral: cuantificación de la epidemia de desinformación Covid-19 en Twitter].

La República (2017) Audio entrevista Juan Carlos Vélez Uribe.

Lesher, M,. Pawelec, H. and Desai, A. (2022) Disentangling Untruths Online: Creators, Spreaders and How to Stop Them, OECD Going Digital Toolkit Notes [Desenredando falsedades en línea: creadores, difusores y cómo detenerlos, OCDE Going Digital Toolkit Notes].

LinkedIn (2023) Government Requests Report [Informe sobre solicitudes gubernamentales].

McIntyre, N., Bradbury, R. and Perrigo, B. (2022) Behind Tiktok's Boom: A Legion Of Traumatised $10-a-Day Content Moderators [Tras el auge de Tiktok: una legión de

moderadores de contenidos traumatizados que ganan 10 dólares al día], Bureau of Investigative Journalism.

Meta (2022) Como aplica Meta sus políticas.

Meta (2022) Meta Q4 2021 Quarterly Update on the Oversight Board [Meta Q4 2021 Actualización trimestral sobre el Consejo Asesor].

Meta (2022) Restricciones de contenido en virtud de la legislación local.

Meta (2023) Contexto local en nuestras normas globales.

Meta (2023) Eliminar contenido infractor.

Meta (2023) Normas comunitarias de Facebook.

Meta (n.d.) Creo que Facebook no tendría que haber eliminado mi publicación.

Ministerio de Tecnologías de la Información y las Comunicaciones (2022) Boletín trimestral del sector TIC: Cifras tercer trimestre de 2022.

Ministerio de Tecnologías de la Información y las Comunicaciones (2022) Índice de Brecha Digital 2021.

Misión de Observación Electoral (MOE) (2018) Impacto de las redes sociales en el proceso electoral colombiano (Elecciones de Congreso y Presidencia 2018).

Misión de Observación Electoral (MOE) (2022) Los discursos de odio racistas y sexistas son legitimadores de la violencia.

Misión de Observación Electoral (MOE) (2022) Sexto informe preelectoral de violencia.

Movilizatorio (2021) Estudio sobre polarización de audiencias en Colombia.

Newton, C. (2021) The Tier List: How Facebook Decides Which Countries Need Protection [La lista de clasificación: cómo decide Facebook qué países necesitan protección].

Observacom (2020) Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta.

Observacom (2021) Declaración latinoamericana sobre transparencia de las plataformas de internet.

Observacom (2022) Moderación privada de contenidos en Internet y su impacto en el periodismo.

Observatorio de Violencias Políticas a las Mujeres (2022) En sus marcas: La carrera de las mujeres en la política.

Osorio, F.E. (2001) Entre la supervivencia y la resistencia. Acciones colectivas de población rural en medio del conflicto armado colombiano.

Pérez, A. and Martinez, C. (2022) Moderación privada de contenidos en Internet y su impacto en el periodismo.

Pinterest (2022) Transparency Report [Informe de transparencia].

Piper, E. (2021) Palestinians Bear the Brunt of Big Tech Moderation [Los palestinos se llevan la peor parte de la moderación de los gigantes tecnológicos].

Privacy International (2018) Privacy and Freedom of Expression in the Age of Artificial Intelligence [Privacidad y libertad de expresión en la era de la inteligencia artificial].

Ranking Digital Rights (2022) Methods and Standards [Métodos y estándares].

Reuters Institute (2022) Digital News Report 2022 Colombia: Consumo y confianza de la información en entornos digitales.

Rotta, S. (2021) Qué pasó con las publicaciones en Instagram durante el paro nacional.

Semana (2016) Consejo de Estado dice que hubo 'engaño generalizado' en campaña del No en el Plebiscito.

Semana (2016) Consejo de Estado podría suspender resultados del Plebiscito.

Silva, S. (2017) Polarización en Colombia: Superar mitos y aceptar realidades, El Eafitense.

Snapchat (2022) Informe de transparencia.

Soyun, A., Baik, S. and Soy, C. (2022) Splintering and Centralizing Platform Governance: How Facebook Adapted Its Content Moderation Practices to the Political and Legal Contexts in the United States, Germany, and South Korea [Escisión y centralización de la gobernanza de plataformas: cómo adaptó Facebook sus prácticas de moderación de contenidos a los contextos políticos y jurídicos de Estados Unidos, Alemania y Corea del Sur], *Information, Communication & Society*, 26, 14: 2843–2862.

Relator Especial de las Naciones Unidas para la Promoción y Protección del Derecho a la Libertad de Opinión y Expresión (2022) Declaración conjunta sobre libertad de expresión y justicia de género.

TikTok (2022) Community Guidelines Enforcement Report [Informe sobre la aplicación de las normas comunitarias].

TikTok (2023) Public Interest Exceptions [Excepciones de interés público].

TikTok (s.f.) Content Violations and Bans [Infracciones y bloqueos de contenidos].

Programa de las Naciones Unidas para el Desarrollo (2021) Internet, libertad de expresión y acceso a la información en Uruguay Aportes para el debate sobre la gobernabilidad democrática en línea.

UNESCO (2021) Hacer frente al discurso de odio en las redes sociales: desafíos contemporáneos.

UNESCO (2021) Declaración de Windhoek + 30: la información como bien común, Día Mundial de la Libertad de Prensa.

UNESCO (2022) Encontrar fondos para que el periodismo prospere : opciones de política para respaldar la viabilidad de los medios de comunicación.

UNESCO (2022) '¿Cómo abordar el discurso de odio en línea con un enfoque basado en los derechos humanos?', video de YouTube.

UNESCO (2022) 'Los límites legítimos a la libertad de expresión: El Plan de Acción de Rabat', video de YouTube.

UNESCO (2022) Transparencia de la moderación privada de contenidos: Una mirada de las propuestas de sociedad civil y legisladores de América Latina.

UNESCO (2023) Salvaguardar la libertad de expresión y el acceso a la información: directrices para un enfoque de múltiples partes interesadas en el contexto de la regulación de las plataformas digitales.

UNESCO (2023) Directrices para la gobernanza de las plataformas digitales.

UNESCO (s.f.) Contrarrestar el discurso de odio.

Asamblea General de las Naciones Unidas (2016) Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión.

Asamblea General de las Naciones Unidas (2018) Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión.

Asamblea General de las Naciones Unidas (2021) La desinformación y la libertad de opinión y de expresión.

Villena, D. (2021) Proyecto de ley pretende regular las redes sociales sin entender cómo funciona Internet.

WFB (s.f.) Country Comparison: Median Age [Comparación entre países: edad media].

X (2022) Colombia.

X (s.f.) Acerca del contenido retenido por el país.

Yann, B. (2017) Claves del rechazo del plebiscito para la paz en Colombia.

Notas finales

[1] Los términos 'discurso de odio' y 'desinformación' no se han definido en el derecho internacional de los derechos humanos. Del mismo modo, aunque muchos responsables políticos y también varias legislaciones se refieren a contenidos 'perjudiciales', no existe un consenso internacional sobre su significado, y carece de una definición jurídica en el derecho internacional de los derechos humanos. Por estas razones, ARTICLE 19 utiliza estos términos entre comillas a lo largo de esta publicación.

[2] La Comisión para el Esclarecimiento de la Verdad, la Convivencia y la No Repetición (CEV) se estableció en el acuerdo de paz de 2016 con las FARC-EP para darle fin al conflicto armado interno que duró más de seis décadas y afectó a más de nueve millones de personas registradas como víctimas. La comisión hizo público su Informe Final el 28 de junio de 2022.

[3] Un informe de la Comisión de la Verdad fue usado como fuente principal para dar una visión general de la historia del conflicto armado en el país.

[4] Esto también ha sido reconocido por la Corte Interamericana de Derechos Humanos (Corte IDH) en los casos *Manuel Isaza Uribe v Colombia* (2018) y *Unión Patriótica (UP) v Colombia* (2023). La Corte declaró que el Estado de Colombia era responsable por violaciones sistemáticas de derechos humanos contra los miembros y militantes del partido político Unión Patriótica (UP) durante más de 20 años. En cuanto al contexto de las violaciones de derechos humanos, la Corte constató que agentes estatales generaron un ambiente de estigmatización contra los miembros de la UP con el fin de excluirlos del juego democrático, presentándolos como 'el brazo político de las FARC'. Este tipo de declaraciones influyeron en la percepción pública, que a su vez influyó en la violencia contra los militantes del partido.

[5] Los hechos controvertidos por las ONGs en la acción de tutela incluían: (1) un discurso pronunciado durante la ceremonia de ascenso de un general del ejército colombiano y retransmitido por la televisión nacional por orden presidencial, en el que el presidente apuntó contra las organizaciones y las calificó de 'escritores y politiqueros que finalmente le sirven al terrorismo y que se escudan cobardemente en la bandera de los derechos humanos', (2) un discurso presidencial tras un atentado en el departamento de Boyacá, en el que el entonces presidente calificó a los defensores de los derechos humanos de

'hablantinosos' sin fundamento, y (3) cuando el entonces presidente, durante su intervención ante la Comisión de Asuntos Exteriores del Parlamento Europeo, calificó al abogado de una ONG colombiana presente en la sala como: 'pertenece a una ONG, El Colectivo de Abogados, que se escuda en detrás de su calidad de organización de derechos humanos para defender a la guerrilla'.

[6] Sentencias T-1191 de 2004 y T-1062 de 2005, y Sentencia T-627 de 2012.

[7] La acción de tutela fue interpuesta por un grupo de más de 1000 mujeres contra el Procurador General de la Nación, la Procuradora Delegada para la Defensa de los Derechos de la Infancia, la Adolescencia y la Familia y la Procuradora Delegada para la Función Pública. Durante tres años, estos funcionarios se pronunciaron en diferentes medios públicos e institucionales, con información inexacta y distorsionada sobre los derechos reproductivos de las mujeres y los métodos anticonceptivos. Entre los hechos que sustentaron la acción de tutela, el Procurador General de la Nación tergiversó una orden de la Corte Constitucional sobre derechos sexuales y reproductivos y publicó un aviso de prensa en el que afirmaba que 'la Corte Constitucional había ordenado implementar campañas masivas de promoción del aborto'.

[8] En la sentencia T-087 de 2021, la Corte estudió una acción de tutela interpuesta por un ciudadano venezolano contra Claudia López, alcaldesa de Bogotá, quien en 2020 se refirió a la situación de seguridad en la ciudad e hizo alusión a la participación de ciudadanos venezolanos en acciones delictivas, señalando expresamente su nacionalidad. La Corte indicó que la alcaldesa excedió su poder-deber de comunicación con los ciudadanos y la libertad de expresión porque la declaración fue discriminatoria. También aclaró que para la situación de seguridad - un asunto de interés público - los funcionarios públicos deben prever los riesgos asociados a sus pronunciamientos, ya que pueden crear o agravar la estigmatización contra ciertos grupos de la población.

[9] En la sentencia T-087 de 2023, la Corte estudió una tutela interpuesta por un grupo de mujeres periodistas sobre la ocurrencia de ataques machistas y misóginos en línea. Estos ataques buscaban invalidar su labor periodística y varios partidos y movimientos políticos se aprovecharon de ello. Las periodistas afirmaron que el Consejo Nacional Electoral (CNE) fue llamado a adoptar medidas para poner fin a este tipo de violencia, pero no lo hizo, lo que alentó la ocurrencia de las agresiones. La Corte concluyó que como los periodistas no informaron al CNE de las agresiones, éste no pudo desplegar las medidas necesarias. No obstante, el fallo reconoció el fenómeno latente de la violencia de género en Internet y sus efectos multidimensionales. La decisión resaltó que las entidades públicas y los partidos políticos deben realizar actividades de prevención y respuesta oportuna ante estas situaciones e indicó que es necesaria una regulación que reconozca y establezca mecanismos específicos para responder a la violencia digital.

[10] Como lo indicó la Comisión de la Verdad: https://www.comisiondelaverdad.co/no-mataras.

[11] La Jurisdicción Especial para la Paz es el mecanismo de justicia transicional colombiano a través del cual los miembros de las FARC, de la Fuerza Pública y tercera que participaron en el conflicto armado son investigados y juzgados.

[12] Véase un informe del *Social Science Research Council* en las campañas de 'desinformación' relacionadas con el plebiscito para la paz de 2016 y las elecciones presidenciales de 2018, que encontró que 'políticos reconocidos son los principales responsables de difundir desinformación en Colombia'.

[13] Como parte de esta campaña, el gobierno vigiló el espacio público digital, y esa información sirvió como base para difundir discursos estigmatizadores y criminalizadores contra los manifestantes. La vigilancia también se utilizó para controlar el espacio físico utilizando la información que recogían para 'anticiparse a los actos vandálicos' y perseguir a las personas.

[14] Fundación para la Libertad de Prensa, (s.f.) *Cartografías de la información*, Bogotá: FLIP.
[15] Citas de una entrevista personal con ColombiaCheck.
[16] El artículo 13 de la Convención prohíbe el 'discurso de odio' entendido como 'toda apología del odio nacional, racial o religioso que constituyan incitaciones a la violencia o cualquier otra acción ilegal similar contra cualquier persona o grupo de personas, por ningún motivo, inclusive los de raza, color, religión, idioma u origen nacional'.
[17] Para una visión general de los marcos regulatorios y estándares relevantes para la moderación de contenidos y la aplicación de los derechos humanos a los mismos, consulte el Manual de moderación de contenidos y libertad de expresión.

[18] Las prácticas de comunicación sobre la aplicación de las políticas varían entre las plataformas digitales. Mientras que los informes de transparencia de algunas plataformas distinguen entre la moderación de contenidos en virtud de las normas comunitarias de la plataforma, las solicitudes legales de retirada y las denuncias de infracciones de la propiedad intelectual, otras dejan claro que el número total de retiradas puede ser el resultado de una combinación de su aplicación de las normas comunitarias y las solicitudes gubernamentales. Algunas plataformas sólo informan del total de contenidos retirados por motivos de aplicación de las normas comunitarias. Este es el caso de Facebook, YouTube y TikTok. Sin embargo, ese indicador no es actualmente trazable en el tiempo. Los informes de transparencia de TikTok, X y Google sólo mencionan el número total de videos eliminados por violar sus lineamientos. Los informes de transparencia de Meta, LinkedIn, Snapchat y Pinterest no contienen información desglosada para Colombia. Los datos sobre moderación de contenidos legales son interesantes: según las restricciones de contenido de Meta basadas en leyes locales, entre julio de 2017 y junio de 2022, la plataforma restringió 1078 piezas de contenido basadas en leyes colombianas en Instagram y Facebook. El informe también especifica que el 90% de las restricciones de contenido debido a solicitudes legales se refieren a publicaciones de Facebook, mientras que en Instagram es aproximadamente 70% contenido y 30% cuentas. Se ofrecen otros detalles,

pero no son suficientes para comprender el alcance de la situación. Para YouTube, los informes de transparencia de Google tienen datos sobre moderación de contenidos derivados de la aplicación de las reglas de la comunidad y de las solicitudes de las autoridades colombianas. YouTube también tiene datos sobre solicitudes de moderación de contenidos por parte de las autoridades locales.

[19] A pesar del gran número de políticas y del mayor esfuerzo por hacerlas más accesibles, su alcance no es del todo comprensible para los usuarios y el número de diferentes políticas que prohíben distintos tipos de contenidos complica las cosas. Para ilustrar este punto, Meta informa de la aplicación de 14 políticas de contenidos en Facebook, 12 en Instagram y 22 políticas en total. X enumera 16 políticas relacionadas con los contenidos. YouTube tiene más de 22 políticas relacionadas con los contenidos. TikTok tiene 26 políticas relacionadas con los contenidos.

[20] Dentro de las definiciones de curación de contenidos y moderación de contenidos utilizadas en este informe, disminución de la visibilidad (downranking) de determinados contenidos (más allá de la eliminación de contenidos) puede, en ocasiones, considerarse simultáneamente una medida de moderación de contenidos y de curación de contenidos.

[21] Cuando los entrevistados hablaron sobre la moderación de contenidos, no especificaron qué acción de moderación de contenidos podía llevarse a cabo (es decir, violación de las normas comunitarias, infracción de los derechos de autor o peticiones estatales). Resulta interesante que al tratar la moderación de contenidos durante la protesta social, como se describe en el estudio de caso 1, la sensación de censura solía estar vinculada a las peticiones de las autoridades.

[22] El informe Guns versus Cellphones explora la sensación de censura que tuvieron los ciudadanos durante la protesta, agravada por otros problemas de moderación de contenidos y falta de transparencia en las plataformas.

[23] La respuesta a una solicitud que Karisma hizo a la Dirección de Derechos de Autor -pidiendo todos los comentarios recibidos durante el trámite del proyecto de ley para reformar la ley de derechos de autor en Colombia en 2018- incluye una carta del Ministerio de Cultura que explicaba el 'concepto de agujero negro' y su petición de incluir el depósito legal digital para que la Biblioteca Nacional aborde el tema. La solicitud del Ministerio que hacía referencia al agujero negro digital tenía como objetivo evitar la pérdida de registros históricos colombianos debido a la naturaleza efímera de Internet y a la incapacidad de preservar las páginas web y el contenido de las redes sociales que contenían dichos registros. La legislación sobre derechos de autor no proporciona el apoyo necesario a la institución.

[24]Este caso resumido se basa en Guns versus Cellphones de la Fundación Karisma e incluye citas de entrevistas realizadas por la FLIP y compartidas con la Fundación Karisma. Se obtuvo el consentimiento de los entrevistados utilizados en este informe.

[25] Testimonio recogido por la FLIP y publicado en Guns versus Cellphones.

[26] Testimonio recogido por la FLIP y publicado en Guns versus Cellphones.

[27] X (2022) *Crisis Misinformation Policy*, consultado el 29 de octubre de 2023.

[28] Por ejemplo, el Consejo Nacional Electoral (CNE) de Colombia firmó un Memorando de Entendimiento con Facebook para las elecciones de octubre de 2019. En México, el Instituto Nacional Electoral firmó un acuerdo similar.

[29] Según los artículos 7 y 8 de la Ley 679 de 2001 y los artículos 5 y 6 del Decreto 1524 de 2002.

[30] Según el artículo 38 de la Ley 643 de 2001.

[31] Algunas autoridades administrativas en Colombia tienen facultades judiciales para bloquear contenidos dando órdenes directamente a los operadores de telecomunicaciones. Esto ocurre en casos de protección de datos, infracciones de propiedad industrial e infracciones de protección al consumidor por parte de la Superintendencia de Industria y Comercio (SIC) según el artículo 54 de la Ley 1480 de 2011, o en casos de infracciones de derechos de autor por parte de la Dirección Nacional de Derechos de Autor.

[32] Según la Ley 1341 de 2009. Durante los estados de emergencia y excepción, el gobierno deberá dictar un decreto específico que indique su alcance, que será revisado por la Corte Constitucional para determinar si las excepciones se ajustan a las normas de derechos humanos.

[33] Se trata de una normativa antigua (en términos de estándares de Internet), de 2001, que utiliza la expresión "pornografía" en lugar de CSAM.

[34] El informe mencionaba acuerdos para combatir contenidos 'ofensivos' (Pakistán) o que 'inciten a la violencia' (Israel). También enumeraba el Código de Conducta de la UE para la Lucha contra la Incitación Ilegal al Odio en Internet, firmado por cuatro empresas principales para eliminar contenidos, comprometiéndose a colaborar con 'trusted flaggers' y a promover 'contranarrativas independientes'.

[35] Este es el caso de YouTube, donde los informes de transparencia de Google contienen datos sobre la moderación de contenidos derivados de las solicitudes de las autoridades locales para la moderación de contenidos en YouTube.

[36] Para el periodo de 2017 a 2019, Meta indica que estas solicitudes responden principalmente a dos temas: (1) artículos que presuntamente violan las leyes relacionadas con la venta de productos regulados y (2) denuncias privadas de difamación. De 2020 a junio de 2022, durante las restricciones de Covid-19, la mayoría de las solicitudes provinieron del Instituto Nacional de Vigilancia de Medicamentos y Alimentos de Colombia (INVIMA), relacionadas con anuncios públicos ilegales sobre productos sanitarios no registrados.

[37] Entre 2011 y junio de 2022, las autoridades colombianas solicitaron a YouTube la retirada de 73 videos. El motivo más común fue la difamación, seguido de la privacidad y la seguridad. La seguridad nacional fue el siguiente motivo más denunciado y se utilizó principalmente entre enero y junio de 2021, durante las protestas sociales. Por último, estaban los motivos de derechos de autor y marcas registradas.

[38] Entrevista con Plurales.

[39] Fue creado por medio del documento CONPES 4080, que contiene la 'Política Pública de Equidad de Género para las Mujeres: Hacia el Desarrollo Sostenible del País'.