Cuadernillo de análisis sobre:

Moderación de contenidos en plataformas de Internet: modelo de gobernanza

Autores:

Alejandro Pisanty Baruch
Issa Luna Pla
Francisco Chan Chan
Jesús Eulises González Mejía
Ernesto Ibarra Sánchez

Coordinadores del proyecto:
Pablo Pruneda Gross / Issa Luna Pla

Contenido

Introducción

El presente ensayo es resultado de un proyecto de colaboración entre el Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México y la Asociación de Internet México (AIMX). El objetivo de la investigación fue explorar el ambiente regulatorio de la moderación de contenidos en Internet, identificar las características que hacen compleja la intervención de actores en la moderación de contenidos y proponer un planteamiento práctico para comprender el alcance de las intervenciones o iniciativas normativas o técnicas por parte de múltiples actores a través de las plataformas, dentro de un marco de afectación controlado por estándares globales de comportamiento en las prácticas de expresión y discursos públicos. Nuestro estudio se dirige principalmente a quienes pueden proponer iniciativas legislativas, regulatorias o judiciales, pero también hacia otros mecanismos como los internos de moderación de las plataformas, o nuevos negocios y modelos.

Entre junio de 2021 y mayo de 2022, el equipo multidisciplinario de investigación revisó fuentes académicas, informes técnicos producidos por organizaciones internacionales, así como políticas, leyes, tratados y jurisprudencia de diversos países, que se complementó con 12 entrevistas a profundidad de expertos[1]. De manera empírica no sistemática, se documentaron actividades, herramientas e iniciativas que intervienen en la identificación de expresiones y contenidos en Internet, que reciben solicitudes de usuarios o cualquier acción que tenga impacto en el control, acceso y limitación del contenido producido o reproducido por las plataformas.

---

* Alejandro Pisanty Baruch es profesor en la Facultad de Química; Issa Luna Pla es investigadora del Instituto de Investigaciones Jurídicas; Jesús Eulises González Mejía es Técnico Académico del Instituto de Investigaciones Jurídicas; Francisco Chan Chan doctorando en derecho en el Instituto de Investigaciones Jurídicas; todos los anteriores de la Universidad Nacional Autónoma de México. Ernesto Ibarra Sánchez es presidente de la Academia Mexicana de Ciberseguridad y derecho digital (AMCID_Mx).
Esta investigación se realizó con fondos derivados de un convenio de colaboración entre la Universidad Nacional Autónoma de México y la Asociación Mexicana de Internet A.C, mismo que resguarda la propiedad intelectual de la obra.

Forma recomendada de citación: Pisanty, Alejandro, Luna Pla, Issa, et al. *Moderación de contenidos en plataformas de Internet: modelo de gobernanza*. Instituto de Investigaciones Jurídicas UNAM - Asociación Mexicana de Internet A.C., junio 2022.

[1] Si bien los hallazgos de esta investigación son exclusivamente responsabilidad de los autores y el equipo redactor, las entrevistas realizadas aportaron y reforzaron ideas y casos de estudio.

Nuestra propuesta consiste en que, para la elaboración de una iniciativa de ley, de tecnología o una propuesta de negocio, entre otros posibles insumos que intervengan en la moderación de contenidos, se apliquen cuatro marcos de referencia en manera sucesiva, lo que llamamos "cribas", para evaluar y en su caso aceptar, rechazar, o mejorar las iniciativas. Si se omite la evaluación que proponemos, es altamente probable que las iniciativas fracasen. Por ejemplo, cuando una ley es inoperante contra algunas conductas en línea que presuponen una fácil detección de los probables responsables de un delito sin considerar que éste puede tener un carácter transnacional y que la identidad de los actores sólo puede ser llevada hasta una dirección IP que para mayor complicación puede ser temporal, falsificada, y no atribuible a ningún individuo en particular, en tanto que se encuentra cometida dentro Internet.

Si bien la idea de cribas es sugerida en las cajas de herramientas (*toolkits*) de ISOC, UNESCO, o estándares de la RELE-CIDH, por mencionar las principales referidas en este estudio, nuestra evaluación de cribas contribuye a argumentar que los actores en el ambiente de Internet (usuarios, organizaciones, compañías y emprendedores, gobiernos) tienen limitados alcances dentro de sus competencias para promover iniciativas de moderación, al tiempo que algunos de ellos tienen espacios reservados para la protección, prevención y resarcimiento. En la línea de UNESCO, elegimos concentrarnos en los temas del proceso de la moderación, en lugar de temas de discursos en disputa y los métodos idóneos.

Complejidad de la regulación del modelo de comunicación en Internet

Internet se ha convertido en el espejo de los mejores y peores aspectos de la humanidad, en el crisol de las más variadas y contrapuestas voluntades, y en un espacio en el que se despliega toda suerte de consecuencias inesperadas de las decisiones. Pasadas las primeras décadas en que construimos y percibimos sus extraordinarios beneficios para la humanidad, cursamos una época en que también se muestran de manera creciente efectos poco deseables y una reacción contra éstos que en ocasiones domina sobre la de los beneficios que continúan creciendo.

Conforme Internet da soporte a cada vez más formatos de información y tipos de actividad, y conforme su uso se extiende a cada vez más sectores de la sociedad, se va convirtiendo en un enorme lente, un prisma en la analogía de Chris Bail, que traduce, a veces amplifica, a veces reduce, y a veces distorsiona la actividad humana. La mayor parte de las variaciones sobre la actividad que ha producido ha sido revolucionaria y positiva, pero también, como todo medio desde que la humanidad aprendió a utilizar la piedra y el fuego, ha servido también para la agresión, el delito, y los más variados desafíos a la convivencia.

Desde sus inicios, ciertamente desde el inicio de su uso por particulares fuera de los gobiernos y la academia, Internet ha requerido intermediarios entre los usuarios y los proveedores, que muchas veces son también los mismos usuarios. Proveedores de conectividad a las telecomunicaciones, proveedores de correo electrónico, de depósitos de archivos, de servicios, y progresivamente con el desarrollo de la tecnología, de alojamiento de servicios, de páginas y portales Web, de plataformas enteras basadas en Web, de servicios de cómputo en la nube, de Internet de las Cosas, y servicios menos visibles a los usuarios finales como los de seguridad, análisis de datos, subastas y publicación de publicidad, intermediarios de pago, y proveedores de estos mismos servicios en la nube, todos son indispensables para el funcionamiento de Internet.

La evolución de la demanda también ha modificado la arquitectura operacional de Internet, en un tránsito de modelos tipo cliente-servidor que hoy nos parecen inocentes, a servicios distribuidos en todo el planeta, integrados con la ubicación y provisión de telecomunicaciones y por otra parte, la evolución de las empresas de telecomunicaciones en la que las antiguas centrales telefónicas se han convertido en centros de datos de cada vez mayor complejidad y en recursos críticos para la vida de las sociedades.

Una de las partes más visibles de Internet, uno de los conjuntos de intermediarios con los que más intensamente interactúan personas e instituciones, son las llamadas “plataformas” en línea, que utilizando la Web como base o al menos como interfase dan lugar a la publicación de contenidos y actividades de los usuarios finales y a la interacción entre los propios usuarios y los contenidos.

Las plataformas, incluyendo no sólo las orientadas a la publicación e interacción sino también los servicios de búsqueda en Internet, software en la nube, comercio electrónico, o servicios dedicados alrededor de los productos de una marca de hardware y software, son objeto de debates y demandas en temas como dominancia, privacidad, monetización de los datos, autonomía y agencia de los usuarios, competencia, poder económico, pago de impuestos, contribución a la democracia o a su debilitamiento, y muchos más. Se han creado a su alrededor temas como "estudios de plataformización" y "capitalismo de vigilancia". Nuestro estudio se mantiene consciente de estas controversias pero se concentra exclusivamente en lo relacionado directamente con la moderación de contenidos.

La publicación de contenidos por parte de usuarios finales individuales que se comunican a través de redes antecede a la difusión de Internet, en los llamados "servicios en línea" y los BBS o "bulletin board systems", así como en listas de correo electrónico y foros en línea basados en tecnologías previas a la difusión de Internet. Si bien desde esos inicios la premisa de la comunicación era la libre expresión de los participantes, se hicieron necesarias algunas reglas y los procedimientos para aplicarlas, con el objeto de evitar que, por ejemplo, se "descarrilaran" las discusiones debido a la polarización, las ofensas y descalificaciones, el uso desproporcionado o la intransigencia, y también debido a la aparición de comunicaciones fuera de tema y posteriormente las comunicaciones comerciales no deseadas o "spam". Vinieron poco después los ataques informáticos que también intensificaron la necesidad de moderar el flujo de las comunicaciones.

La moderación de los debates y las conversaciones en las redes se ha basado frecuentemente en la autorregulación de las comunidades. Algunas comunidades virtuales particularmente longevas, como la Computational Chemistry List que existe desde 1991, han sobrevivido y prosperado a lo largo de décadas con base en acuerdos tomados por sus propios miembros, sin necesidad de autoridad externa ni de mecanismos coercitivos internos particularmente fuertes.

Al expandirse las interacciones a las actuales redes sociales en línea y otros espacios de conversación como los juegos en línea, los sitios de comercio electrónico, los servicios de contacto para la formación de parejas, y muchos más, la escala de las interacciones ha crecido de manera desbordante. Estas interacciones se producen entre personas que pueden no conocerse ni convivir físicamente sino en el "espacio de los flujos" del sociólogo Manuel Castells, entre personas de países, culturas, historias personales, ocupaciones, ideologías, idiomas, y contextos dispares y desconocidos unos para los otros. Las interacciones pueden pasar rápidamente de la conversación civil y moderada a la hostilidad, y ésta volverse organizada, transmisible al espacio físico a nivel de poner en riesgo vidas, e incluso amenazar la seguridad de las sociedades. Los contenidos, más allá del texto, ahora sonido, video, imagen, imágenes sintéticas, conductas programadas mediante algoritmos, identidades artificiales, y software, pueden poner en riesgo derechos tan elementales como la vida misma.

Los responsables de estas conductas son siempre humanos o sus productos, instituciones, algoritmos, asociaciones, gobiernos, empresas. Los intermediarios se encuentran muchas veces ante dilemas escabrosos, con contenidos y actividades que en algunas sociedades o partes de ellas son inaceptables, pero sin haberlos producido, seleccionado, vigilado, filtrado, autorizado, o respaldado, a razón de millones de unidades por segundo. A diferencia de un medio de difusión tradicional como una estación de radio o televisión, o de la prensa, el intermediario puede elegir – y generalmente lo hace - no ejercer control editorial sobre la actividad de los usuarios. Puede, quizás, intervenir mediante un motor de recomendación y otros algoritmos en la selección de los contenidos que se presentan ante cada uno de los usuarios, generalmente con base en intereses propios de la plataforma como puede ser producir *engagement* (enganchar) para incrementar el valor de la publicidad que es la base de su modelo de negocios, pero en general en estas condiciones no interviene como responsable o corresponsable de la publicación.

En atención a esta circunstancia, y en particular en Estados Unidos, se estableció una ley y una jurisprudencia que exime a los intermediarios de responsabilidad civil sobre los contenidos emitidos por los usuarios, siempre que el intermediario no tenga intervención en seleccionar lo que se presenta, esté dispuesto a eliminar contenidos que sean declarados ilegales o cuando tenga conocimiento de otras infracciones, y actúe de buena fe.

La moderación de contenidos empieza, para algunos autores, en la recomendación, pero generalmente se refiere a la selección de contenidos ya publicados que deben ser eliminados, o bien al menos no deben ser difundidos ante algunos usuarios. Para determinar qué es aceptable como contenido difundido en las redes, cada plataforma establece un conjunto de reglas y de procedimientos para modificarlas y aplicarlas. En algunos casos las reglas son prácticamente un “todo se vale”, en otros, las reglas son explícitas en cuanto a determinados contenidos – palabras, imágenes – y pueden ser fijadas por el intermediario o bien ser desarrolladas, total o parcialmente, por la comunidad de usuarios. Además pueden existir leyes nacionales aplicables, por ejemplo para contenidos que una sociedad determinada haya declarado ilegales o contrarios a la convivencia, y acuerdos internacionales como los contrarios a la trata de personas y al abuso sexual contra menores y personas sin capacidad plena de decisión.

Cada día se producen miles de millones de publicaciones y cada día una parte de éstas, millones, es objetada por algún motivo, legítimo o abusivo (es decir, acusaciones en falso con intención de agredir o acallar al emisor). Cada una de estas objeciones debe ser procesada por un programa de computadora, una persona física, o una combinación de ambas según la plataforma. De manera típica, las personas que participan en esta función tienen unos segundos para procesar cada objeción y darle paso o negarla. En casos particulares en los que el contenido observado viola la ley o ciertas normas, el caso debe ser elevado a una persona de mayor jerarquía en la empresa y en caso de violar la ley, se debe informar a las autoridades para que procedan de acuerdo a derecho en la persecución de un delito.

La escala de esta operación es mayúscula, frecuentemente se ejerce a través de personal tercerizado y distribuido en el mundo para afinar la sensibilidad lingüística y cultural, y cuenta con el apoyo de sistemas automatizados sofisticados, de alto rendimiento, y en constante evolución. Sin embargo, la magnitud de estas operaciones están por debajo de la escala de las publicaciones e incluso de las objeciones, y puede estar en una escala de tiempo y velocidad también insuficiente. Por ejemplo, cuando un criminal utiliza una red social para transmitir en vivo sus acciones, los segundos o minutos que pasan entre que empieza a hacerlo y es detectado, y se puede suprimir su transmisión, pueden haber permitido que el hecho se propague mundialmente de manera irreversible mediante copias y retransmisiones también en otras redes.

El problema de la moderación de contenidos se complica con el hecho de que en la mayoría de los países muchas expresiones en ellos son legales, a pesar de que puedan ser objetables por muchas otras razones por una parte de la sociedad. Las diferencias de criterio entre lo que es aceptable o no son intensas, y no lo es menos la opinión sobre la compatibilidad entre estos contenidos y las libertades de expresión, acceso a la información, y asociación lícita. El problema se vuelve aún más complejo cuando las expresiones son vistas sin comprensión de su contexto – una ironía inocente puede ser leída como una grave ofensa o una incitación a la violencia – en un entorno internacional e intercultural.

Se han intentado múltiples formas de intervención de los gobiernos en el mundo para tratar de corregir o al menos amortiguar aquello que incomoda a la sociedad o algunos de sus representantes. En algunos países se ejerce un control férreo sobre el tránsito internacional de la información *ex-ante* y sobre lo que dice cada ciudadano *ex post*, hasta el punto de pretender controlar también lo que piensa y siente. Otras sociedades privilegian la libertad de expresión y recomiendan que contra la expresión que incomode u ofenda se responda con más expresión, más ideas y puntos de vista, para un debate público que haga progresar a la sociedad.

En el intento de corregir problemas reales o percibidos se corre el riesgo de modular las conductas de tal manera que se pierdan aspectos esenciales de Internet, como su interoperabilidad y apertura. Este riesgo reconocido desde hace décadas también ha dado lugar a algunos de los análisis que contiene nuestro estudio.

Los discursos en Internet, cuando menos, desafían a los sistemas jurídicos constitucionales e internacionales en la multiplicidad de las fuentes de interpretación de derecho (que suceden dentro de la versatilidad de las formas de la moderación), en la identificación de los autores de las expresiones dañinas, en la extra-territorialidad de los actos indebidos, en la ausencia de una rectoría jurídica exclusiva por parte de los Estados.

La moderación es relevante para los actores del entortno en tanto que protege a los usuarios de contenido no deseado, ofensivo o dañino, resguarda su privacidad y seguridad en el medio, pretende detener la dispersión de mensajes de odio, de incitación a la violencia, de discriminación, noticias o información falsa, e identifica delitos como abuso sexual contra menores, la pornografía infantil, las extorsiones y fraudes. Igualmente, fomenta la confianza de los usuarios en las plataformas y en la tecnología, e incentiva el comercio electrónico y el uso de tecnologías para comunicarse a través de las fronteras. En ningún otro medio como en Internet, la libertad de expresión y el libre flujo de las ideas es facilitado por la innovación tecnológica, y por ello, frecuentemente la moderación es una práctica impulsada por los acuerdos multilaterales de comercio, como es el caso del Tratado entre México, Estados Unidos y Canadá (T-MEC).

Gobernanza de la moderación en Internet

Los actores que intervienen a través de acciones de moderación en el entorno son múltiples y actúan de manera individual o coordinada, lo que obedece a la muy reconocida cualidad de gobernanza multisectorial de Internet, incluso protegida por los sistemas globales de derechos humanos. Se ubican al menos tres actores principales: Estados (y sus instituciones), corporaciones y empresas, y organizaciones de la sociedad civil; un cuarto actor que conviene diferenciar es la comunidad técnica cuyo expertise e incidencia determinan el comportamiento de las plataformas. Estos emprenden estrategias de control de contenidos o defensa de derechos en lo individual o coordinadamente.

Los Estados intervienen a través de regulaciones directas y marcos constitucionales con potencial impacto en la moderación, y a través de las resoluciones de los poderes judiciales en los que se litigan y reconocen derechos y se resarce daños, que generan precedentes para normar el comportamiento en ciertos territorios (con impacto persuasivo en el comportamiento extraterritorial de las plataformas).

Algunos ejemplos importantes de formas de autorregulación y regulación nacional sobre moderación de contenido en plataformas digitales son el *Code of conduct on countering illegal hate speech online*, de la Comisión Europea en acuerdo con Facebook, Microsoft, Twitter y YouTube; el *Netwotk Enforcement Act* (NetzDG-2017) ley de Alemania, que obliga a las plataformas con más de dos millones de usuarios a proveer a los usuarios mecanismos para que puedan notificar a la plataforma cuando existe contenido considerado indebido, y establece plazos para que las plataformas remuevan contenido “ilegal obvio”; o bien, el *Online Safety Act* (OSA) (2021) en Australia que solicita a la industria el desarrollo nuevos códigos para regular el contenido ilegal y restringido. En cuanto a los casos judiciales, una selección de sentencias y casos en moderación de contenidos puede encontrarse en el banco de jurisprudencia de la iniciativa Global Freedom of Expression de la Universidad de Columbia.

Las corporaciones intervienen en los contenidos en Internet a través de sus estrategias y prácticas humanas o automatizadas, formalizadas o experimentales, transparentes y no transparentes de cara a los usuarios, y políticas internas y herramientas tecnológicas, por su cuenta o a través de proveedores de servicios externos. Las políticas y estrategias de autorregulación alineadas a los objetivos de la moderación se enfocan en eliminar el contenido no permitido, y detener la escalabilidad de los discursos dañinos, incluso los basados en desinformación e ideologías extremas. En términos generales la literatura concuerda en que estas herramientas son humanas (instrumentadas por personas), automatizadas (por máquinas o algoritmos) o mixtas que se aplican *ex-ante* y *ex-post* de la publicación de contenidos. Múltiples plataformas remueven contenido *automatizadamente ex-ante* (como spam o pornografía infantil) y atienden millones de reportes *ex-post* sobre contenido que probablemente viola las normas comunitarias de la plataforma o derechos humanos. Adicionalmente, podrían existir esquemas como el Consejo Asesor de Facebook que emite recomendaciones a la compañía iniciadas por un mecanismo de apelación y de detección proactiva, decide con base en las reglas de la plataforma, el derecho internacional de los derechos humanos y el test tripartito reconocido en el derecho internacional de la libertad de expresión.

Las organizaciones no gubernamentales intervienen en la moderación de contenidos a través de la promoción, monitoreo y movilización de usuarios para impulsar normas técnicas y/o de conducta, y promueven el conocimiento de los derechos de los usuarios en las redes. Éstas promueven la exigencia para que las corporaciones y empresas difundan el número de publicaciones eliminadas y cuentas suspendidas derivadas de sus políticas o lineamientos, notificar a los usuarios cuando su contenido o sus cuentas son suspendidas, y permitirles la oportunidad de apelar las decisiones, por ejemplo, a través de las recomendaciones en el informe 2021 de la organización Artículo 19.

Los actores del Estado y las empresas adoptan códigos y estrategias permanentes o temporales conjuntas; las empresas y las organizaciones mantienen diálogos y acciones coordinadas a través de foros globales, y los tres sectores se unen en temas urgentes comunes como el combate al terrorismo y la violencia. En el curso de los años estas acciones han producido impacto, cuando menos, en la priorización de los temas globales de la moderación de contenidos, en las propuestas para crear medidas preventivas y resarcitorias, y en la creación de grupos de expertos con temas de frontera. Ejemplos significativos son los Principios de Manila sobre la responsabilidad de los intermediarios, los Principios de Santa Clara sobre transparencia y rendición de cuentas en la moderación de contenidos, el Plan de Acción de Rabat sobre la libertad de expresión contra la incitación al odio promovido por la Oficina del Alto Comisionado de la Organización de las Naciones Unidas.

Evaluación de iniciativas por medio de cribas

La moderación de contenidos puede ser filtrada por estas cuatro cribas conforme evolucione. Cada iniciativa de ley o regulación que incida en la moderación de contenidos puede ser evaluada en términos de la escalabilidad de la medida en comparación con la escala de la conducta que se quiere modular; de la identidad, o bien anonimato y pseudonimias asociadas; los efectos del tráfico y operación a través de múltiples fronteras nacionales o subnacionales sobre la definición de la conducta, su aceptabilidad y legalidad, y sobre la posibilidad de ejercer acción penal mediante autoridades de un tercer país (o varios), así como el hecho de que la mayoría de los países, incluido México, no tiene autoridad en la jurisdicción en que están establecidas las plataformas y la mayoría de los proveedores de servicios que las hacen funcionar; el abatimiento de barreras, generalmente deseable para la formación de organizaciones y acceso a mercados pero indeseable para la facilidad de formación de organizaciones delictivas; la reducción de fricción que favorece la comunicación humana, el comercio, la educación, y el acceso a la salud, pero también favorece la inmediatez que aprovechan los delincuentes; y la memoria paradójicamente inmensa e indeleble pero a la vez frágil y manipulable de Internet.

Aplicadas en orden, las cuatro cribas constituyen una herramienta invaluable que reúne cajas de herramientas existentes. Sociedad, Legislativo, Ejecutivo, industria y los innovadores encontrarán mediante ella rutas seguras hacia una mejor Internet, pues permite ubicar el alcance de las intervenciones de los actores, es una técnica para anticipar impactos no deseados en derechos humanos, en los principios y la infraestructura de Internet y, finalmente, comunica la complejidad del entorno.

1. Características esenciales de Internet

El riesgo de producir un deterioro de Internet, hasta que pierda su esencia incluso, es latente y creciente a pesar de la probada robustez de la red. Para prevenir este riesgo, la Internet Society (ISOC) ha desarrollado un catálogo de "invariantes de Internet" que deben ser preservadas; una descripción de "la manera Internet de conectar" y una caja de herramientas analíticas para producir una "manifestación de impacto en Internet", similar a las de impacto regulatorio o ambiental vigentes en otros ámbitos, que permite detectar cuándo una medida puede tener un efecto deletéreo sobre Internet. La "caja de herramientas" de ISOC incluye las siguientes:

1. Una infraestructura accesible con un protocolo común
2. Una arquitectura abierta de componentes básicos interoperables y reutilizables
3. Gestión descentralizada y un único sistema de enrutamiento distribuido
4. Identificadores globales comunes
5. Una red de uso general y neutralidad tecnológica

La infraestructura de Internet se puede fragmentar por las intervenciones de filtrado o bloqueo y generar partes no accesibles o bien por la utilización de protocolos que no son compartidos universalmente (de uso común). La arquitectura abierta es violada cuando se introduce el uso obligatorio de componentes que no son interoperables, como software o el paso obligatorio del tráfico por puntos de control ("gatekeepers") con tecnología propietaria que pudiera ser utilizada en intervenciones de moderación o bloqueo de contenidos.

La gestión descentralizada puede ser violada mediante la imposición de un sistema de gestión de redes centralizado o mandatado, el uso de "gatekeepers" que imponen una centralización para fines políticos o comerciales de control de contenidos, o bien mediante la imposición de esquemas de enrutamiento como los que en algunos casos se proponen para la implementación de leyes y políticas de "Internet soberana". Sobre los identificadores globales comunes, éstos pueden transgredir mediante esquemas "nacionales" que utilizan direcciones IP propias (redundantes con las globales, pero detrás de artificios como NATs), versiones del sistema de nombres de dominio "nacionales", y otros similares.

Finalmente, la red de uso general y neutralidad tecnológica es violada cuando se restringen los usos posibles de la red, por ejemplo prohibiendo su utilización para servicios de video en línea o de mensajería instantánea, o cuando se impone por ley o regulación una tecnología específica o se prohíben otras. Esto es uno de los efectos deletéreos de algunas regulaciones sobre criptografía propuestas en algunos países. Una colección de análisis de casos se encuentra en https://www.internetsociety.org/issues/internet-way-of-networking/internet-impact-assessment-toolkit/ e incluye legislación de países tan diversos como Canadá, el Reino Unido, Nigeria y Camboya.

2. Seis factores de la infraestructura de Internet

Para comprender mejor la relación entre, por un lado, conductas y motivaciones de larga data en la humanidad, y por otro, los efectos disruptivos y revolucionarios de Internet, Alejandro Pisanty ha creado un sistema de seis factores de la infraestructura de Internet que permite hacer un análisis de una iniciativa de ley o de negocio, identificar si es o no compatible con Internet, y en algunos casos identificar una dirección de mejora hacia esta compatibilidad. El sistema ha sido usado ya en la construcción de capacidades en el Senado mexicano, en el análisis de estrategias nacionales de ciberseguridad, y en el análisis de la respuesta de Internet a la pandemia de COVID-19 y sus efectos sobre una Internet abierta y libre, así como a las normas de convivencia entre Estados en el ciberespacio.

1. *Escala* - ¿los factores críticos de la iniciativa tienen escalabilidad conmensurable con la "escala Internet"? Si el esquema de moderación depende de procesos manuales, por ejemplo, no será escalable.

2. *Identidad* - ¿qué hipótesis explícitas o tácitas sustentan a la iniciativa en lo que hace a identidad, anonimato, pseudónimos, y su administración? Por ejemplo, si para aplicar una ley persiguiendo un delito es necesario asegurar la identidad de los responsables en una cadena comisoria compleja, es altamente probable que la ley termine siendo letra muerta al no poder llevar a cabo los procesos forenses y procedimientos civiles o penales que lleven a identificar y capturar a los responsables de la conducta. También fallará si se ancla la identidad en direcciones IP o nombres de dominio.

3. *Transjurisdiccional* - ¿la ley o tecnología de moderación de contenidos propuesta responde adecuadamente a la conducta que atraviesa fronteras? (generalmente nos referimos a fronteras nacionales pero las diferencias jurisdiccionales pueden presentarse también en el nivel subnacional). Si la persecución del delito enunciado por la ley depende de la colaboración entre autoridades de diversos países, corre el riesgo de convertirse en inoperable toda vez que en uno o más de ellos no será posible obtener la colaboración deseada, o bien, la conducta no será considerada delictiva en alguno de los sistemas jurídicos domésticos.

4. *Abatimiento de barreras* - ¿la propuesta depende de la erección de barreras al acceso a recursos en línea, a barreras de entrada a mercados, a barreras para la formación de organizaciones virtuales formales o informales? Por ejemplo, si la iniciativa de moderación legislativa o técnica está orientada a individuos o grupos organizados actuando de manera delictiva, la medida fracasará ante la fluidez con la que se forman, transforman y desaparecen dichas organizaciones criminales en el ciberespacio. Igualmente, si se erigen barreras artificiales al comercio, Internet las rodeará a través de mecanismos que posiblemente favorezcan a nuestros competidores.

5. *Reducción de fricción* - ¿es la fricción un enemigo o un aliado de la parte a la que la iniciativa se propone beneficiar? Las iniciativas que reducen la fricción favorecen el comercio, la agilidad y con ello, la adopción de trámites gubernamentales, la colaboración y la educación. En cambio, en algunas ocasiones es necesario aumentar la fricción, por ejemplo, para permitir que las personas reflexionen y así eviten caer en trampas que las convierten en víctimas de delitos (como sucede en *phishing*).

6. *Memoria*. ¿La iniciativa crea memoria o la destruye? ¿Qué usos buenos y malos se pueden dar a los acervos de información? ¿Cuándo se decreta una forma de "olvido", se borra efectivamente la memoria, o solamente se dificulta el acceso para algunas partes y es facilitado para otras? La iniciativa se modificará para, por ejemplo, reducir la captura y conservación de información, reducir o controlar el acceso de terceros a esta, o bien, asegurar que la información que de todas maneras será pública esté disponible en forma inmediata y accesible desde un principio.

3. Valores globales de libertad de expresión en Internet

Las instituciones internacionales de derechos humanos del sistema universal y los sistemas regionales reconocen valores comunes de la protección a la libertad de expresión en Internet, así como reglas para la intervención y no intervención de los gobiernos y de las plataformas en sus respectivos ámbitos regulativos. Desde las mismas instancias se han producido documentos cada vez más dirigidos a las complejidades en Internet, que resumen principios y reglas de las libertades de expresión e información dirigidos a legisladores, jueces, moderadores de contenidos en plataformas y sociedad civil, principalmente, la caja de herramientas para actores judiciales de la UNESCO y los Estándares para una Internet libre, abierta e incluyente de la RELE-CIDH y el Plan Acción de Rabat promovido por la Oficina del Alto Comisionado en Derechos Humanos de las Naciones Unidas. Las herramientas de interpretación de la libertad expresión, se comprenden desde de los cuatro principios D-A-A-M de la universalidad de Internet

acuñados por la UNESCO, y sus respectivos indicadores que promueven el vínculo entre Internet y los derechos, éstos son:

D- que internet está basada en los Derechos humanos

A- que sea abierta

A-que sea accesible para todos

M-que cuente con la participación de múltiples partes interesadas

La presente síntesis no pretende simplificar la extensa interpretación de las instituciones internacionales y nacionales de derechos humanos, tampoco excluir la casuística en la interpretación que deben llevar a cabo los actores, para la que recomendamos usar bancos de jurisprudencia y los instrumentos internacionales. Las iniciativas de los múltiples actores moderadores de contenidos en cualquiera de sus formas y niveles se alinean a los valores cuando:

1. Protegen las dos dimensiones de la libertad de expresión: la libertad individual y la libertad de la sociedad de expresarse y recibir e investigar información, así como su función indispensable para la democracia.

2. Introducen alguno de los límites admitidos a la libertad de expresión: regulados generalmente por el Estado por medio de una ley o interpretación judicial, para lo que se requiere aplicar un test tripartito. Los principales límites admitidos son las leyes de difamación y de protección de los derechos de terceros, como la privacidad, la honra y reputación.

3. Pasa por un test tripartito para limitar la libertad de expresión: que consiste en: 1) que la limitación a la libertad de expresión sea definida en forma precisa y clara a través de una ley formal y material y orientada al logro de objetivos del derecho internacional de derechos humanos; 2) que la limitación sea necesaria e idónea en una sociedad democrática; y 3) que sea estrictamente proporcional a la finalidad perseguida.

4. Que respete y proteja los derechos al honor, reputación o privacidad: Cuando se trata de proteger estos derechos, además del test tripartito, el estándar global implica determinar un daño probable o inminente a estos derechos, y la aplicación de un test de necesidad estricta para probar que la medida de restricción es absolutamente indispensable.

5. Permita el libre flujo de información y expresiones y no censura previa, bloqueos o eliminación de sitios de internet: fuera de los contenidos estrictamente prohibidos *ex ante* de su publicación (como contenidos de abuso de niños, niñas o adolescentes o propiedad intelectual), la regla es que los actores que pretenden intervenir en la moderación de contenidos permitan el libre flujo de contenidos y promuevan medidas preventivas.

6. Respeten y protejan el libre pensamiento y no la censura indirecta: formas de limitar la libertad de expresión indirectamente son aquellas que silencian, inhiben o crean un efecto de autocensura en el discurso, a nivel individual o colectivo.

7. Moderan y contienen los discursos no protegidos por la libertad de expresión: en muchos países prohibidos por las leyes, como el discurso de odio, incitación a la discriminación, hostilidad o violencia y pornografía infantil. Un instrumento relevante para identificar estos discursos es el Plan Acción de Rabat.

8. Introduzca esquemas para combatir la publicación de noticias falsas y la desinformación y mitigue el escalamiento.

9. La no discriminación de personas en el tratamiento de los datos y en el tráfico y acceso, que pueda además estar basada en la autoría, el contenido, el destino del contenido, o el equipo tecnológico desde el que se pueda acceder.

10. Transparente las prácticas, métodos y datos sobre los esquemas de moderación, bloqueo o filtrado de contenidos.

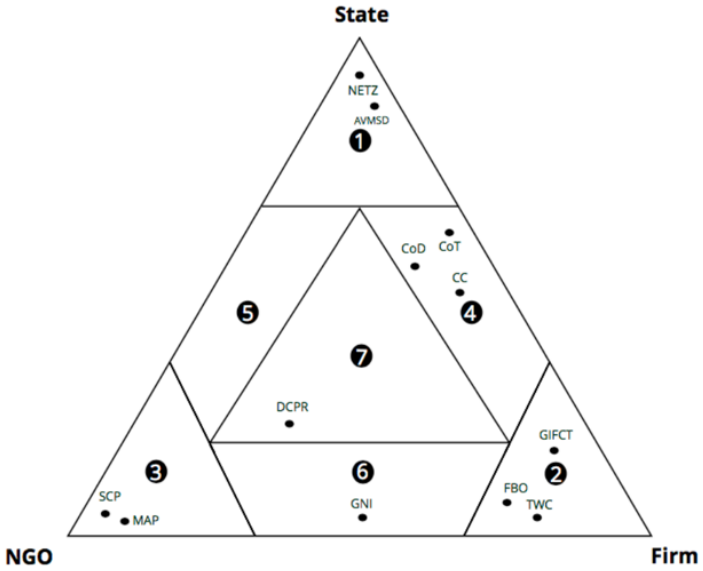
11. Que introduzca reglas del debido proceso en los procesos y esquemas de moderación de contenidos para garantizar derechos de las personas autoras de contenidos, propietarios de sitios web, y derechos de usuarios y las audiencias.

12. Que no imponga responsabilidades legales a los proveedores de servicios de internet por los contenidos producidos por terceras personas, a menos que sea a través de una corte que resuelva en contra de contenido que es considerado ilegal en aras de proteger a los usuarios de daños causados por los mismos.

3. Triángulo de Gorwa

La gobernanza de Internet es un sistema descentralizado, democrático, enfocado en la resolución de problemas, flexible y adaptable, en el que participan todos los sectores – academia, industria, gobierno, sociedad civil, y comunidad técnica. Con base en el probado éxito histórico de organizaciones y mecanismos como ICANN o el Anti-Phishing Working Group y el Internet Governance Forum, Robert Gorwa ha adaptado el "triángulo de gobernanza" de Abbot y Snidal, diseñado para organizaciones internacionales multisectoriales, a las organizaciones relacionadas con Internet.

Este triángulo permite clasificar las iniciativas y mecanismos de moderación de acuerdo a su nivel de formalidad (alcance normativo), la distribución de la toma de decisiones entre los actores, el número de actores que deben proponer la iniciativa dependiendo del proceso de moderación desde gobierno (regulación legal), industria (autorregulación) y sociedad civil (producción de normas e influencia, o participación) y multiactores. El modelo permite anticipar la competencia por la legitimidad en el ambiente de la regulación que jugará cada iniciativa frente a otras existentes y las relaciones de poder.



1. Regulación directa por un estado nación (p.ej. la legislación “NetzDG” de Alemania).

2. Autorregulación corporativa (p.ej. el “Content Oversight Board” o Consejo de Supervisión de Contenido de Facebook; aunque podría disputarse que es en parte un modelo de co-regulación con otros actores).

3. Gobernanza y monitoreo por la sociedad civil organizada (p.ej. Principios de Santa Clara para la Moderación de Contenidos, Principios de Manila sobre Responsabilidad de Intermediarios)

4. Co-gobernanza entre estado y empresa (p.ej. ACCC News Media and Digital Platforms Mandatory Bargaining Code” o “Código obligatorio de negociación entre medios noticiosos y plataformas digitales” de la Comisión Australiana de Competencia y el Consumidor.

5. Co-gobernanza entre Estado y sociedad civil organizada

6. Co-gobernanza entre empresa y sociedad civil organizada (p.ej. GNI “Global Network Initiative”, “Contract for the Web”)

7. Co-gobernanza entre Estado, empresa, y sociedad civil organizada (p.ej. "Christchurch Call", "Convocatoria de Christchurch contra el Terrorismo")

Recomendaciones para los actores y retos en la materia

Primero. – En los últimos años, derivado de la gobernanza multi-actor de Internet, se reconoce que las iniciativas y esquemas de moderación de contenidos deben alinearse con los principios generales de la red y su arquitectura, con los factores de la infraestructura y los derechos humanos, especialmente, con los valores globales de la libertad de expresión. Especialmente en el contexto del capítulo 19 del T-MEC, cualquiera que sea la iniciativa de intervención y del sector del que provenga, requerirá necesariamente de ser valorada en función del rol de los actores, sus competencias y el alcance de sus decisiones pasando al menos por las cribas que este estudio propone.

Segundo. – Estamos conscientes de las tensiones entre los actores, la competencia y la necesidad de la legitimidad de las medidas e iniciativas, de las molestias expresadas por actores dentro de diversos sectores a causa del fenómeno social, tecnológico y económico de los flujos y discursos en las plataformas en Internet. Al mismo tiempo, valoramos el diálogo multisectorial, el desarrollo de la innovación tecnológica y la necesidad de promover la gobernanza de la regulación en la moderación de contenidos a través de cribas y estándares y mediante un trabajo colaborativo (multistakeholder). La implementación de cualquier marco regulatorio debe estar geográficamente limitado y en concordancia con el derecho internacional de los derechos humanos.

Tercero. – Es apremiante una discusión profusa con especialistas en labores de moderación de contenidos que permitan identificar las problemáticas comunes a las plataformas por motivos de diferencias lógicas, lingüísticas, informativas, semióticas, históricas, socio-culturales y jurídicas del contexto en los discursos y su moderación. Esta discusión debe incluir a los procesos y tiempos de decisión y derivar en protocolos de actuación para moderadores y desarrolladores de esquemas tecnológicos, basados en prácticas estandarizadas y de actuación prioritaria y urgente, cuando los daños no pueden ser evitados todos y oportunamente.

Fuentes consultadas

Abbott, K. W., & Snidal, D. *The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State.* In W. Mattli & N. Woods (Eds.), *The Politics of Global Regulation* (pp. 44–88). Princeton, NJ: Princeton University Press. 2009. Doi: 10.1515/9781400830732.44

Bail, Chris. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press, 2021.

B. Solum, Lawrence, "Models of Internet governance", Bygrave, Lee A. y Bing Jon (eds.), *Internet Governance. Infrastructure and Institutions*, New York, Oxford University Press, 2009.

BID, ALAI, Responsabilidad de intermediarios de Internet en América Latina: Hacia una regulación inteligente de la economía digital. Disponible en *https://publications.iadb.org/publications/spanish/document/Responsabilidad-de-intermediarios-de-Internet-en-Am%C3%A9rica-Latina-Hacia-una-regulacion-inteligente-de-la-econom%C3%ADa-digital.pdf*

Castells, Manuel. *La era de la información*. Vol.1: La sociedad red. 2da Ed. Alianza, Madrid, 1996.

CIDH, Relatoria Especial para la Libertad de Expresión. Estándares para una Internet libre, abierta e incluyente. Capítulo III del Informe Anual 2016 de la Relatoría Especial, 15 de marzo 2017.

Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Las plataformas en línea y el mercado único digital Retos y oportunidades para Europa. *https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52016DC0288&from=EN*

Eric Schmidt y Jared Cohen, *The new digital age: transforming nations, businesses, and our lives*. Vintage Books, A Division of Random House LLC, New York, First Vintage Books Edition, 2014.

Gillespie, Tarleton, *Custodians of the Internet*, New Haven & London, Yale University Press, 2018.

Gorwa, R., *The platform governance triangle: conceptualizing the informal regulation of online content,* Internet Policy Review, 2019.

Grimmelmann, James, "The Virtues of Moderation", *Yale Journal of Law and Technology,* EEUU, 2015.

Internet Society Global Internet Report, Consolidation in the Internet Economy, 2019.

Klonik, Kate, "Why the History of content moderation matters", Techdirt, 2018. Disponible en: https://perma.cc/3V7A-SDZ9.

Kummer, Marcus, La gobernanza en Internet: Vayamos al grano, Unión Internacional de Telecomunicaciones, http://www.itu.int/itunews/manager/display.asp?lang=es&year=2004&issue=06&ipage=governance

Kurbalija, Jovan, Gobernanza en Internet: Asuntos, actores y brechas, KGP, 2005. https://books.google.com.mx/books?hl=en&lr=&id=uhkbxgmEWu4C&oi=fnd&pg=PA8&dq=info:SRAUwwN8jAgJ:scholar.google.com&ots=sh_96YXJQf&sig=wpePkPn69EOmhGW9tmgLZXkPaoM&redir_esc=y#v=onepage&q&f=false

Lanza, Edison y Matías, Jackson, Moderación de Contenidos y Mecanismos de Autorregulación. "El Oversight Board" de Facebook y sus implicancias para América Latina, El Diálogo, 2021.

Perset, K. "The Economic and Social Role of Internet Intermediaries", *OECD Digital Economy Papers*, No. 171, OECD Publishing, Paris, 2010.

Pírková, Eliška y Pallero, Javier, *26 recomendaciones sobre gobernanza de contenido. Una guía para legisladores, reguladores y encargados de políticas empresariales*, Access Now.

Pisanty Baruch, Alejandro. "The vexing problems of oversight and stewardship in Internet governance" en "The Working Group on Internet Governance" https://www.apc.org/sites/default/files/IG_10_Final_0.pdf

Pisanty Baruch, Alejandro. «Gobernanza de Internet y los Principios Multistakeholder de la Cumbre Mundial de la Sociedad De La Información», Revista Mexicana De Política Exterior, n.º 79-80 (marzo):9-39. https://revistadigital.sre.gob.mx/index.php/rmpe/article/view/643.

Pisanty, Alejandro, Open Internet Governance: The 6F Framework And COVID-19, 2020, https://www.medianama.com/2020/05/223-open-Internet-governance-6f-framework/

Privacy International, Article 19, "Privacy and Freedom of Expression in the Age of Artificial Intelligence", 2018. Available at: https://www.Article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf

Radu, Roxana, *Negotiating Internet Governance*, NY, Oxford University Press, 2019.

Riedl, J. M, *et. al.*, "Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States", *Policy and Internet*, Wiley Periodicals, 2020.

Rosen, Jeffrey, *Google Gatekeepers*, NYT, 2008.

Schweich, Barbara van, *Internet Architecture and Innovation*, England, The MIT Press, 2010.

Unión Europea. Power structures in content moderation, "The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act", 2021.

UNESCO, *Global toolkit for judicial actors. International legal standards on freedom of expression, access to information and safety of journalists,* 2021.

Documentos Internacionales

Agenda de Túnez para la Sociedad de la Información, WSIS-05/TUNIS/DOC/6(Rev.1)-S, https://www.itu.int/net/wsis/docs2/tunis/off/6rev1-es.html

CIDH, Relatoría Especial para la Libertad de Expresión, Libertad de expresión e Internet, OEA/Ser.L/V/II CIDH/RELE/INF.11/13, 31 diciembre 2013.

Declaración Conjunta sobre la Independencia y la Diversidad de los Medios de Comunicación en la Era Digital del Relator Especial de las Naciones Unidas sobre la promoción y protección del derecho a la libertad de opinión y de expresión, el Representante para la Libertad de los Medios de Comunicación de la Organización para la Seguridad y Cooperación en Europa (OSCE), el Relator Especial para la Libertad de Expresión de la Organización de los Estados Americanos (OEA) y el Relator Especial sobre Libertad de Expresión y Acceso a la Información de la Comisión Africana sobre Derechos Humanos y de los Pueblos (CADHP), 2018. Disponible en: https://www.oas.org/es/cidh/expresion/showarticle.asp?artID=1100&IID=2.

Declaración Conjunta sobre libertad de expresión en Internet del Relator Especial de las Naciones Unidas (ONU) para la Libertad de Opinión y de Expresión y la Relatora Especial para la Libertad de Expresión de la Comisión de Derechos Humanos de la OEA, Washington, D.C., 20 de enero de 2012. Disponible en: https://www.oas.org/es/cidh/expresion/showarticle.asp?artID=888&lID=2

European Court of Hu-man Rights, Guide on Article 8 of the Convention – Right to respect for private and family life, agosto 2021.

Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, A/HRC/35/22, 30 de marzo de 2017.

ONU, El derecho a la privacidad en la era digital, A/RES/68/167. 21 de enero de 2014.

ONU, La desinformación y la libertad de opinión y de expresión, Informe de la Relatora Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, Irene Khan, A/HRC/47/25, 13 de abril de 2021.

ONU. Declaración Conjunta sobre libertad de expresión en Internet del Relator Especial de las Naciones Unidas para la Libertad de Opinión y de Expresión y la Relatora Especial para la Libertad de Expresión de la Comisión de Derechos Humanos de la OEA, Washington, D.C., 20 de enero de 2012. Disponible en: https://www.oas.org/es/cidh/expresion/showarticle.asp?artID=888&lID=2

Relatoría Especial para la Libertad de Expresión, “Capitulo IV (Discurso de odio y la incitación a la violencia contra las personas lesbianas, gays, bisexuales, trans e intersex en América)” en Informe Anual 2015, Informe de la Relatoría Especial para la Libertad de Expresión, OEA/Ser.L/V/II, Doc. 48/15, 31 de diciembre de 2015.

Casos

Corte IDH, Caso Granier y otros (Radio Caracas Televisión) Vs. Venezuela. Excepciones Preliminares, Fondo, Reparaciones y Costas. Sentencia de 22 de junio de 2015. Serie C No. 293.

Corte IDH. Caso López Álvarez Vs. Honduras. Fondo, Reparaciones y Costas. Sentencia de 1 de febrero de 2006. Serie C No. 141.

Corte IDH. Caso Fontevecchia y D`Amico Vs. Argentina. Fondo, Reparaciones y Costas. Sentencia de 29 de noviembre de 2011. Serie C No. 238,.

Corte IDH. Caso Herrera Ulloa Vs. Costa Rica. Excepciones Preliminares, Fondo, Reparaciones y Costas. Sentencia de 2 de julio de 2004. Serie C No. 107.

Corte IDH. Caso Fontevecchia y D'Amico Vs. Argentina. Fondo, Reparaciones y Costas. Sentencia de 29 de noviembre de 2011. Serie C No. 238., párr. 53; y CIDH, Informe Anual 2010, Informe de la Relatoria Especial para la libertad de Expresión, Volumen II, OEA/Ser.L/V/II. Doc. 69, 30 diciembre 2011, Capítulo III, párr. 347.

Corte IDH. Caso Herrera Ulloa Vs. Costa Rica. Excepciones Preliminares, Fondo, Reparaciones y Costas. Sentencia de 2 de julio de 2004. Serie C No. 107.

Corte IDH. Caso Kimel Vs. Argentina. Fondo, Reparaciones y Costas. Sentencia de 2 de mayo de 2008. Serie C No. 177, párr. 83.

Corte IDH. Caso Kimel Vs. Argentina. Fondo, Reparaciones y Costas. Sentencia de 2 de mayo de 2008. Serie C No. 177, párr. 83.

Amparo Directo en Revisión 4865/2018, https://www.scjn.gob.mx/sites/default/files/listas/documento_dos/2019-10/ADR-4865-2018-191009_0.pdf

TEDH, Beizaras y Levickas Vs. Lithuania, application no. 41288/15, 14 de enero de 2020.

TEDH, Sanchez Vs. Francia, application no. 45581/15, 2 de septiembre de 2021.

TEDH, Ürçdaq vs. Turquía, application no. 23314/19, 31 de agosto de 2021.

Personas expertas entrevistadas

Avina, Marissa; Capelo, Maria Cristina; Careaga, Javier; Centeno, Danya; García Urbina, Daniel; García, Luis Fernando; Maldonado, Leopoldo; Piña, Carlos; Ríos, Agustín; Sosa Pastrana, Fernando; Trejo Delarbre, Raúl; y Valles, María Andrea.