CONTRIBUCIÓN DE INTERNETLAB A EXPEDIENTE T-8.764.298 ACCIÓN DE TUTELA INSTAURADA POR ESPERANZA GÓMEZ SILVA CONTRA FACEBOOK COLOMBIA S.A.S, INSTAGRAM COLOMBIA Y META PLATFORMS, INC, MINISTERIO DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES (MINTIC) Y SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO.

Francisco Brito Cruz, director ejecutivo de InternetLab
Fernanda Martins, directora de InternetLab
Clarice Tavares, coordinadora del área de Desigualdades e Identidades de InternetLab
Iná Jost, coordinadora del área de Libertad de Expresión de InternetLab

1. ¿Cómo funciona la inteligencia artificial en la moderación de contenidos en plataformas de redes sociales?

Antes de abordar el uso y funcionamiento de la inteligencia artificial en la moderación de contenido, es necesario dar un paso atrás y hacer una breve apreciación sobre qué es la moderación de contenido en sí[1]. Consideramos que esta es **una actividad de elaboración y aplicación de reglas**. Con el fin de construir diferentes ambientes digitales propicios para las interacciones sociales y la generación de contenido por parte de sus usuarios, cada plataforma decide qué tipos de discurso (es decir, contenido en texto o multimedia) serán permitidos, motivados, desalentados o prohibidos. Para ello, las redes sociales elaboran '*guidelines*', '*standards*' u otros conjuntos de reglas que se adjuntan a sus términos de servicio disponibles públicamente. Internamente, por su parte, detallan procedimientos para interpretar estas reglas privadas, organizan precedentes y construyen sistemas de aplicación para manejar todo el contenido generado por sus usuarios.

Todas estas actividades, por lo tanto, forman parte del mencionado binomio de funciones: elaborar y aplicar las normas que rigen y gestionan el comportamiento de los usuarios en un espacio determinado. Administran, por lo tanto, la expresión de estas

[1]"La moderación de contenido consiste en un proceso mediante el cual las plataformas de Internet actúan sobre cuentas o contenidos que violan sus términos de uso, afectando su disponibilidad, visibilidad y/o credibilidad. La moderación puede implicar diferentes medidas, como la eliminación, la suspensión temporal, la reducción artificial del alcance o prominencia, la superposición de una pantalla de advertencia, la adición de información complementaria, entre otras". [Thiago Dias Oliva, Victor Pavarin Tavares y Mariana G Valente, "Una solución única para toda la internet? Riesgos del debate regulatorio brasileño para la operación de plataformas de conocimiento", Diagnósticos y Recomendaciones (São Paulo: InternetLab, 2020)].

personas que, en el caso de las redes sociales más populares, generan cantidades inimaginables de publicaciones diariamente[2].

Así, la apreciación que se debe hacer sobre el uso de la automatización en la moderación de contenido se refiere al enorme **desafío logístico impuesto a los administradores por la cantidad de contenido que circula en las redes sociales**. Cientos de millones de 'piezas' de expresión se publican diariamente en todo el mundo, en diferentes idiomas y en coyunturas sociales, políticas, culturales y jurídicas diversas. Entendemos que es imposible prescindir de sistemas automatizados para detectar, recopilar y organizar estas publicaciones y actividades en línea, así como para establecer mecanismos efectivos de aplicación de reglas privadas de conducta, sin incurrir al menos en enormes costos e inversiones. La ausencia de estas herramientas podría, de hecho, hacer logísticamente prohibitiva la tarea de moderar el contenido en las grandes plataformas digitales.

Un desafío adicional es la velocidad con la que se dispersan los contenidos y cambian los contextos, y por ende, los significados de lo que potencialmente puede ser considerado hostil, lo que configura otra dificultad para la moderación. Errores en estas tareas pueden ser graves y llevar a la supresión de discursos de diferentes grupos sociales. Cuando estos discursos son emitidos por personas pertenecientes a poblaciones históricamente marginadas, podemos observar el efecto de fortalecimiento de su invisibilización y silenciamiento social que también les afecta fuera de las redes sociales.

El campo de la inteligencia artificial[3] se suma a este esfuerzo al hacer posible 'enseñar' a las computadoras a reconocer patrones y señales típicas en el contenido publicado por los usuarios, interviniendo especialmente en tareas como la detección, priorización y análisis de contenidos que violan las reglas establecidas por las plataformas. Con esto no queremos decir que se deba prescindir de las revisiones humanas, sino más bien que sería imposible depender únicamente de individuos para moderar las actividades de los

[2]Datos de la edición más reciente de la investigación '*Data Never Sleeps*' muestran, por ejemplo, que cada minuto se suben 500 horas de video en YouTube y se publican más de 347 mil tweets. Disponible en: https://www.domo.com/data-never-sleeps#.

[3] De acuerdo con Bianca Kremer y Ramon Costa, "la inteligencia artificial como la conocemos hoy en día se basa principalmente en el aprendizaje automático (*machine learning*). Esto significa que estas aplicaciones basadas en IA están desarrolladas de manera que las máquinas puedan aprender a crear modelos y obtener resultados sobre problemas que deseamos resolver. La capacidad de estas tecnologías es exponencial, lo que permite que entreguen muchos resultados de manera ágil y en grandes volúmenes." [Costa, R. S., Kremer, B. Inteligência artificial e discriminação: desafios e perspectivas para a proteção de grupos vulneráveis frente às tecnologias de reconhecimento facial. Revista Brasileira de Direitos Fundamentais & Justiça, 16 (1). Disponible en: https://dfj.emnuvens.com.br/dfj/article/view/1316/1065].

usuarios en las plataformas, una tarea demasiado 'artesanal'. Además de hacer inviable el servicio, podría generarse un escenario en el que las plataformas incurran en más fallas, ya que sería imposible moderar activamente todo el contenido publicado en las redes sociales. Por lo tanto, consideramos que una de las buenas prácticas en el sector es encontrar un equilibrio y complementariedad entre sistemas automatizados (y 'inteligentes') y la participación de colaboradores humanos que supervisen, revisen y orienten el trabajo de las máquinas.

La inteligencia artificial entra, por lo tanto, como un artefacto fundamental en la moderación a gran escala. Automáticamente, identifica expresiones potencialmente infractoras y las compara con los parámetros de la comunidad. Calcula, contabilizando los flujos de su propio aprendizaje, si el contenido viola las reglas y decide qué debe ser retirado o mantenerse en línea.

En resumen, las políticas y términos de uso de las plataformas se implementan a través de sistemas que pueden contener elementos de inteligencia artificial. Estos sistemas permiten, entre otras actividades, mantener publicaciones legítimas y eliminar publicaciones dañinas que infringen las reglas. Finalmente, es fundamental concluir este capítulo con dos premisas de nuestras actividades de investigación:

(i) Toda inteligencia artificial desarrollada para esta función está propensa a cometer errores en sus tareas de moderación de contenido, actuando frecuentemente con esquemas que indican la probabilidad de que cierto contenido pueda ser infractor;
(ii) Existen formas de mitigar los errores o sesgos cometidos por la inteligencia artificial, pero no de eliminarlos por completo, ya que estos provienen de realidades cambiantes, complejas y en constante movimiento, inherentes a la dinámica de la comunicación humana en sociedades altamente conectadas (este punto se profundizará en la pregunta 2).

2. ¿Existe evidencia sobre posibles sesgos en las herramientas de inteligencia artificial empleadas para adelantar actividades de moderación de contenidos en redes sociales? Si conoce dicha evidencia, por favor compártala con esta Sala, especialmente sobre sesgos contra las mujeres que desarrollan actividades sexuales pagadas offline o fuera de la plataforma.

edifício itália
avenida são luís 50
1º andar conjunto 11b
1046-000 são paulo sp brasil
www.internetlab.org.br

Como explicamos anteriormente, cualquier herramienta de inteligencia artificial puede cometer errores y mostrar sesgos no deseados.[45] Hemos observado el desarrollo de investigaciones que, además de demostrar las posibles reproducciones de desigualdades estructurales, también señalan la potencialización de violencias que algunos mecanismos de inteligencia artificial pueden ofrecer. En este sentido, es importante enfatizar que las tecnologías, incluida la inteligencia artificial, no son neutrales, ya que son producidas y entrenadas por personas que llevan consigo sus propias experiencias y perspectivas, y tampoco funcionan de manera estática.

Uno de los problemas está relacionado con las bases de datos utilizadas para el entrenamiento de *machine learning*. A menudo, las bases de datos utilizadas para entrenar el algoritmo para ejecutar ciertas tareas tienen limitaciones - pueden ser incompletas o tener sesgos -, y también pueden presentar problemas de actualización - los datos que alimentan la inteligencia artificial no serán válidos para siempre. Además, la inteligencia artificial aprende patrones a partir del comportamiento de los usuarios, lo que puede generar la reproducción de sesgos no deseados no solo del banco de datos inicial. Por más sofisticadas y probadas que sean, con bases de datos complejas y diversas, las tecnologías de inteligencia artificial también pueden cometer errores al encontrarse con nuevas realidades, contextos y problemas que no estaban bien representados en sus bases de datos de entrenamiento.

Como era de esperarse, dado que toda inteligencia artificial está sujeta a errores y es imposible eliminar por completo su riesgo, tenemos evidencia de la detección de sesgos en sistemas de moderación de contenido que utilizan inteligencia artificial en temas relacionados con la sexualidad. Estas evidencias muestran las dificultades específicas enfrentadas por las plataformas que tienen reglas privadas aplicables a estos contenidos, constituyendo casos relevantes para reflexionar sobre las estrategias implementadas por las empresas de redes sociales para mitigar problemas en su moderación de contenido.

1. Una investigación realizada por InternetLab probó la herramienta Perspective, una tecnología desarrollada por Jigsaw (del grupo Alphabet) que mide el nivel de toxicidad de los contenidos en texto, comparando cómo la tecnología clasificaba los escritos publicados por *drag queens* y por nacionalistas blancos. La

[4]Para obtener más información sobre el sesgo algorítmico, consulte: NOBLE, Safiya Umola. Algorithms of oppression: how search engines reinforce racism. Nueva York: Ney York University Press, 2018.

[5] SILVA, Tarcízio. Racismo Algorítmico em Plataformas Digitais: microagressões e discriminação em código. In: Anais do IV Simpósio Internacional LAVITS – Assimetrias e (In)visibilidades: Vigilância, Gênero e Raça. Salvador, Bahia, Brasil, 2019.

investigación señala que este tipo de tecnología puede no ser capaz de distinguir matices de contextos, como por ejemplo, la diferencia entre discursos de odio contra personas LGBTQIA+ y contenido publicado por personas LGBTQIA+ que a menudo reinterpretan términos considerados ofensivos, dándoles un significado positivo para la comunicación entre pares. Los resultados indicaron que los perfiles de drag queens fueron considerados potencialmente más tóxicos que los perfiles de supremacistas blancos. La herramienta de inteligencia artificial consideraba potencialmente tóxicos términos comúnmente utilizados por personas LGBTQIA+, como '*sissy*', '*gay*', '*lesbian*' y '*queer*', independientemente del contexto en que se utilizaran.[6] El caso de la investigación sobre las *drag queens* demuestra distorsiones y sesgos de la inteligencia artificial, que no fue capaz de identificar las sutilezas específicas de la comunicación de personas *queer*, en la que el uso de un lenguaje 'pseudo-ofensivo' es una forma común de comunicación entre miembros de la propia comunidad.[7] Además, este ejemplo sugiere que las bases de datos utilizadas para entrenar la inteligencia artificial eran inconsistentes o parciales, dado que el algoritmo interpretaba como tóxicos términos que, en sí mismos, no son necesariamente violentos u ofensivos.

2. Otro ejemplo que ilustra las limitaciones de las tecnologías de inteligencia artificial se refiere a los casos de moderación imprecisa y/o equivocada en publicaciones que involucran desnudez. Algunas plataformas prohíben, de manera legítima, la difusión de imágenes que contienen partes del cuerpo desnudo, como es el caso de Meta. Esto ocurre no solo por decisiones de negocio sobre qué tipo de contenido erótico es o no aceptable, sino también porque una elección como esta puede reducir los riesgos de circulación de contenidos de explotación sexual de niños y adolescentes debido a la dificultad para determinar la edad precisa en imágenes y videos.[8] En general, cuando hay previsión de prohibición de este tipo de contenido, se establecen como excepciones a estas políticas imágenes con desnudez que hagan referencia a (i) momentos de parto y

[6] Dias Oliva, T., Antonialli, D.M. & Gomes, A. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. Sexuality & Culture 25, 700–732 (2021). https://doi.org/10.1007/s12119-020-09790-w; and InternetLab. "Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet?". 28 de junho de 2019. Disponible en: https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/.

[7] Murray, Stephen O. "The Art of Gay Insulting." Anthropological Linguistics, vol. 21, no. 5, 1979, pp. 211–23. JSTOR, http://www.jstor.org/stable/30027635.

[8] En respuesta al Comité de Supervisión de Facebook, Meta afirmó, por ejemplo, que al elaborar los principios generales de sus políticas sobre desnudez, consideró: (1) la naturaleza privada o sensible de las imágenes; (2) si se dio consentimiento para obtener y compartir imágenes de desnudez; (3) el riesgo de explotación sexual; y (4) si la divulgación de las imágenes podría llevar a acoso fuera de la plataforma, especialmente en países donde podrían ser culturalmente ofensivas. Disponible en: https://oversightboard.com/decision/BUN-IH313ZHJ/.

posparto, (ii) procedimientos quirúrgicos y médicos de confirmación de género, y (iii) autoexámenes de cáncer o publicaciones sobre prevención y evaluación de enfermedades en genitales y partes íntimas.[9] Sin embargo, existen casos recurrentes de eliminación de contenido, especialmente de personas trans y no binarias, con publicaciones sobre procedimientos de afirmación de género, basadas en las políticas de desnudez y de contenido de índole sexual. En respuesta a estas eliminaciones, personas trans y no binarias se unieron al movimiento #DeserveToBeHere (en español, #MerecemosEstarAqui), en el que cuestionaban la prohibición de fotos con desnudez sobre la transición de género.[10]

El tema de la identidad de género y la desnudez fue objeto de cuestionamiento en el ámbito del Comité de Supervisión de Facebook[11], al cual InternetLab envió un comentario. En el caso, una pareja de personas trans cuestionaba la eliminación de fotos en las que mostraban sus senos expuestos y con los pezones cubiertos con cinta adhesiva, como parte de una campaña para recaudar fondos para una cirugía de afirmación de género.[12] En su decisión, el propio Comité revocó la decisión de Meta de eliminar las fotos y subrayó la necesidad de revisar y mejorar las orientaciones sobre moderación de contenido, como una forma de reducir errores en eliminaciones.[13] Por lo tanto, es esencial mejorar el análisis del contexto para la moderación de contenido, especialmente en casos que involucren a grupos históricamente marginados, que son afectados de manera desproporcionada.

Sin embargo, como se mencionó anteriormente, es importante destacar que, a pesar de los casos de errores y discriminación, así como las limitaciones, el uso de inteligencia

---

[9]Política de desnudez adulta y actividades sexuales de Meta. Disponible en: https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/.

[10] Disponible en: https://www.thepinknews.com/2021/04/22/instagram-trans-bodies-censorhip-we-deserve-to-be-here/

[11] El Comité de Supervisión de Facebook es un órgano independiente responsable de revisar las decisiones de moderación de contenido en Facebook e Instagram y hacer recomendaciones sobre las políticas de moderación de contenido en la plataforma. En su sitio web, el Comité afirma que su objetivo es 'promover la libertad de expresión a través de decisiones independientes basadas en principios con respecto al contenido en Facebook e Instagram y mediante la emisión de recomendaciones sobre la política de contenido relevante de la empresa de Facebook'. Disponible en: https://www.oversightboard.com/.

[12]Para obtener más información, consulte la contribución de InternetLab al caso 2022-009-IG-UA & 2022-010-IG-UA del Comité de Supervisión de Facebook. MARTINS, Fernanda; TAVARES, Clarice; FERREIRA, Lux. 'Comentario de InternetLab sobre el caso 2022-009-IG-UA & 2022-010-IG-UA del Comité de Supervisión de Facebook'. Disponible en: https://internetlab.org.br/pt/noticias/internetlab-envia-contribuicao-para-caso-de-identidade-de-genero-e-nudez-do-comite-de-supervisao-do-facebook/.

[13] La decisión del Comité de Supervisión de Facebook está disponible en: https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/.

artificial para la moderación de contenido en plataformas de redes sociales es inevitable. El volumen de publicaciones, la diversidad de idiomas y contextos hace que sea imposible que la aplicación de las reglas de las plataformas sea realizada exclusivamente por humanos. Según los datos de la investigación *DataNeverSleep* de 2020, llevada a cabo por Domo,[14] a cada minuto se publican 347 mil nuevas historias en Instagram y se suben 147 mil fotos en Facebook.[15] Por lo tanto, en un contexto en el que el uso de herramientas de inteligencia artificial es imprescindible debido al desafío logístico de moderar miles de contenidos publicados por minuto, es necesario que estas herramientas estén en constante mejora. Deben contar con bases de datos más completas y complejas, prestando atención detenida a los contextos y a las especificidades locales y de grupos históricamente marginados. Además, como se desarrollará a continuación, la inteligencia artificial debe combinarse con otros recursos para la moderación de contenido, ya que la máquina por sí sola, en su estado actual de desarrollo, no es capaz de detectar y comprender todas las particularidades de la comunicación humana.

3. ¿Qué estrategias pueden implementarse para superar los posibles sesgos de las herramientas de inteligencia artificial empleadas en la moderación de contenidos en redes sociales? Si conoce evidencia al respecto, por favor compártala con esta Corporación judicial.

Dado que se asume que el ejercicio de la moderación de contenido implica el uso de inteligencia artificial, así como que estas herramientas generan sesgos que impactan negativamente en la gestión de la expresión en línea, pasamos a hablar sobre los mecanismos que pueden ayudar a mitigar estas consecuencias.

La primera vertiente de acciones para atenuar estos riesgos se refiere a **estrategias de transparencia sobre los mecanismos de moderación, combinadas con métodos de rendición de cuentas y reglas sobre el debido proceso en la moderación de contenido.** Para empezar, es fundamental que las normas y términos de uso que rigen las plataformas estén escritos en un lenguaje claro y accesible, traducidos a los idiomas de todas las regiones donde se navegan, y dispuestos de manera que cualquier persona pueda encontrarlos. También, que existan métodos de rendición de cuentas, es decir,

[14] Domo es una empresa de software en la nube que publica anualmente el DataNeverSleep, un estudio en el que presenta el número de contenidos publicados en cada plataforma por minuto. Disponible en: https://www.domo.com/.

[15] TechTudo. "O que acontece a cada minuto na Internet? Estudo traz dados surpreendentes". 14 de agosto de 2020. Disponible en: https://www.techtudo.com.br/noticias/2020/08/o-que-acontece-a-cada-minuto-na-internet-estudo-traz-dados-surpreendentes.ghtml. Los datos completos de la investigación de Domo están disponibles en: https://www.domo.com/data-never-sleeps#.

que las personas que acceden a las redes sociales puedan dar opiniones y *feedbacks* sobre el ejercicio de la moderación. Esta herramienta tiene el potencial de generar un ciclo virtuoso de mejora, ya que la inteligencia artificial se nutre de su propia fuente, mejorándose con el uso. De esta manera, la aplicación de los términos de uso y el rendimiento del sistema estarían siendo constantemente evaluados. Por último, que las plataformas tengan instancias de apelación, de manera que las personas puedan, a través de procedimientos claros y formales, conocer la razón de las decisiones que influyen en su expresión en las redes, y cuestionarlas si lo consideran justo y necesario.

El segundo ámbito de acciones fundamentales para atenuar errores de la automatización es la incorporación de etapas de supervisión realizadas por personas calificadas que revisen y dirijan el trabajo de las máquinas. Aunque estamos de acuerdo con el hecho de que la automatización a través del uso de inteligencia artificial es indispensable, consideramos que estos procesos deben complementarse con recursos humanos que los mejoren y agreguen matices a sus análisis, frenos y contrapesos para mitigar posibles sesgos. Esta revisión añade tonos a las decisiones maquinadas, sumando especificidades y contextos que pueden ser piezas clave para una gestión saludable del discurso que circula en las plataformas. Esto es especialmente importante cuando la posibilidad de cometer errores afecta desproporcionadamente a los usuarios, como en casos de poblaciones históricamente minorizadas cuyo discurso ya sufre limitaciones fuera del ambiente en línea. InternetLab aborda esta cuestión en una investigación que profundiza el estudio de sistemas de moderación de contenido en capas[16].

Por último, se señala que sistemas como estos necesitan inversión y reevaluación constantes. Ante la enorme cantidad de contenido, su velocidad de circulación y los frecuentes cambios de contexto que involucran la expresión, es necesario que las plataformas dediquen una mirada cuidadosa y continua a esta cuestión. Esto debe involucrar, por ejemplo, diálogo con especialistas locales, equipos interdisciplinarios con puntos focales regionales atentos y especializados, capaces de abordar desafíos relacionados con idiomas y coyunturas específicas, como un proceso electoral, la promulgación de cierta legislación o el inicio de un conflicto. Además, es fundamental que existan métricas de análisis específicas y transparentes, y que todos estos mecanismos combinados sean capaces de impulsar mejoras en las herramientas tecnológicas y, por consecuencia, en la gestión de la moderación de contenidos.

[16] La moderación en capas trae etapas adicionales de análisis calificado para ciertos tipos de perfiles o contenidos. Esta herramienta puede ser una estrategia empleada por las empresas para mitigar riesgos a los derechos humanos, ya que dan prioridad de análisis a algunos tipos de usuarios o publicaciones que deben ser cuidadosamente revisados para proteger ciertos tipos de discurso, como por ejemplo, expresiones de poblaciones históricamente marginadas, activistas políticas, periodistas, entre otros.
El estudio también está disponible en su versión en inglés.

4. Por último, según su conocimiento del funcionamiento de las plataformas de red social ¿cómo sus operadores previenen convertirse en intermediarios o gatekeepers de otras plataformas o sitios web con contenido pornográfico? Por ejemplo, frente al uso de enlaces a plataformas con contenido para adultos explícito como sitios web de pornografía o redes sociales propias del negocio webcam (ej. OnlyFans)

InternetLab no tiene investigaciones específicas sobre el papel de intermediarios o "*gatekeepers*" ejercido por plataformas de redes sociales en relación al mercado de producción de contenido adulto o al ecosistema de distribución de pornografía lícita en línea. Sin embargo, sobre este tema, hemos formulado al menos dos puntos que deben ser considerados para una reflexión acerca del papel de las plataformas como intermediarios de contenido pornográfico.

El primer punto es que, en el caso del contenido adulto, existen capas adicionales de sensibilidad debido al riesgo de que actividades extremadamente dañinas se aprovechen de lagunas y permisos, como en la producción y difusión de contenido de explotación sexual, lo que puede incluir no solo a personas adultas, sino también a niños y adolescentes, o en la divulgación no consentida de imágenes íntimas.

Un segundo punto a destacar es que **las plataformas de redes sociales son, por excelencia, intermediarias de contenidos de terceros**. Es decir, no son autoras de los contenidos que se publican en las redes sociales, sino que alojan contenidos de otros individuos. Como intermediarias, las plataformas tienen la responsabilidad de establecer reglas a través de sus políticas internas para determinar qué actividades son legítimas o ilegítimas, basándose en el modelo de servicio en línea que pretenden ofrecer. La capacidad para desarrollar diferentes reglas, con diversas formas de moderación de contenido, es clave para que los espacios creados en línea tengan vocaciones propias, creando un internet plural y diverso, en el que cada una de las plataformas pueda tener una identidad propia, con usuarios e intereses específicos.[17]

Esta diversidad de modelos de servicio puede ser ejemplificada con las diferentes formas en que las plataformas abordan las políticas sobre desnudez. Mientras que algunas plataformas como X, antiguo Twitter, permiten la desnudez y los usuarios

[17] Artur Pericles Lima Monteiro, Francisco Brito Cruz, Juliana Fonteles da Silveira e Mariana G. Valente, "Armadilhas e caminhos na regulação da moderação de conteúdo", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponible en: https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pd.

pueden esperar tener acceso a este tipo de contenido, otras, como Facebook e Instagram, prohíben este tipo de contenido como parte de las características de los entornos en línea que hacen posible su modelo de negocio. Además, existen casos como Tumblr, una red social utilizada especialmente para la divulgación de contenidos artísticos, que modificó sus políticas sobre desnudez. En 2018, Tumblr dejó de permitir contenido sexual explícito o que contenga desnudez para evitar la eliminación de su aplicación de la App Store, y para adaptarse mejor a las políticas de los procesadores de pagos y servidores.[18] Más recientemente, la plataforma volvió a permitir solo contenidos con desnudez, excluyendo actividades sexuales.[19] Las modificaciones en las reglas de Tumblr demuestran que estas plataformas pueden cambiar sus políticas con el tiempo y que hay una variedad de relaciones a considerar entre diferentes tipos de intermediarios, no solo entre plataformas de redes sociales y sitios web.

La multiplicidad de reglas y prácticas de moderación en lo que respecta a contenido sexual y desnudez dinamiza la producción y difusión de contenido en diferentes plataformas, permitiendo a los usuarios una diversidad de formas de relación y uso de los servicios de redes sociales. En una plataforma que permite contenido con desnudez, el usuario tiene la expectativa de encontrar una cierta expresión, más orientada al público adulto, mientras que en aquellas donde este tipo de contenido está prohibido, se espera que el tipo de interacción y lenguaje sean apropiados para todas las edades.[20] **Por lo tanto, cada plataforma tiene espacio y libertad para diseñar sus políticas de manera diferente, como una forma de reforzar la libertad de expresión, ya que esta multiplicidad permite la creación de espacios que albergan expresiones diversas, fomentando la pluralidad de discursos**.

[18] The Verge. “Tumblr will now allow nudity but not explicit sex”. Disponible en: : https://www.theverge.com/2022/11/1/23435516/tumblr-porn-ban-modified-nudity-allowed-sexually-explicit-depictions-banned.

[19]En su política actual, la plataforma afirma que "la desnudez y otros tipos de contenido adulto en general son bienvenidos. No estamos aquí para juzgar tu arte, solo pedimos que agregues una advertencia de contenido para que las personas puedan optar por filtrarlo del panel, si lo prefieren". Disponible en: https://www.tumblr.com/policy/br/community.

[20] Artur Pericles Lima Monteiro, Francisco Brito Cruz, Juliana Fonteles da Silveira e Mariana G. Valente, “Armadilhas e caminhos na regulação da moderação de conteúdo”, Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponible en: https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pd.