Opinión de asesoramiento normativo sobre el programa de verificación cruzada de Meta

I. Resumen ejecutivo

En octubre de 2021, tras publicarse en el Wall Street Journal información sobre el programa de verificación cruzada de Meta, el Consejo asesor de contenido aceptó una solicitud de la empresa para revisar dicho sistema y ofrecer recomendaciones para mejorarlo. Esta opinión de asesoramiento normativo es nuestra respuesta a esa solicitud. Analizamos el sistema de verificación cruzada de Meta teniendo en cuenta los compromisos de la empresa con los derechos humanos y sus valores declarados, lo que nos hizo plantearnos cuestiones importantes acerca de cómo la empresa trata a sus usuarios más influyentes. Cuando el Consejo empezó a trabajar en esta opinión de asesoramiento normativo, Meta manifestó que, en aquel momento, llevaba a cabo aproximadamente 100 millones de intentos diarios de aplicar sus políticas. Con semejante volumen, aunque la empresa pudiera tomar decisiones sobre el contenido con un 99% de precisión, seguiría cometiendo un millón de errores cada día. En este sentido, aunque un sistema de revisión de contenido debería tratar equitativamente a todos los usuarios, el programa de verificación cruzada se enfrenta a desafíos mayores al moderar volúmenes de contenido tan grandes.

Según Meta, tomar decisiones a esa escala implica que, en ocasiones, se elimine erróneamente contenido que no infringe sus políticas. Para evitar que esto suceda, el programa de verificación cruzada recurre a fases adicionales de revisión manual cuando se trata de determinadas publicaciones que, en una primera instancia, se habían identificado como infractoras. Cuando los usuarios que figuran en las listas de verificación cruzada de Meta publican contenido de estas características, este no se elimina de inmediato, como es el caso para la mayoría de las personas, sino que permanece publicado a la espera de revisión manual adicional. Meta denomina este tipo de verificación cruzada "revisión secundaria de respuesta temprana". A finales de 2021, la empresa incluyó en el proceso de verificación cruzada determinadas publicaciones marcadas para revisión adicional en función del contenido, y no de la identidad del autor. Meta denomina este tipo de verificación cruzada "revisión secundaria general".

Al llevar a cabo nuestro análisis, detectamos varias deficiencias en el programa de verificación cruzada de Meta. La empresa había manifestado al Consejo que uno de los fines del programa de verificación cruzada es promover sus compromisos con los derechos humanos. Sin embargo, consideramos que dicho programa parece orientado más bien a satisfacer los intereses de Meta. El Consejo es consciente de que se trata de una empresa,

pero el hecho de ofrecer protección adicional a ciertos usuarios (elegidos en gran medida en función de intereses empresariales) hace posible que contenido que, en condiciones normales, se eliminaría rápidamente permanezca publicado durante más tiempo, lo que puede provocar daños. Asimismo, creemos que Meta no realizó un seguimiento de los datos para saber si la verificación cruzada se traduce en una toma de decisiones más adecuadas. Por otra parte, también manifestamos nuestra preocupación por la falta de transparencia del programa.

Teniendo en cuenta todo lo anterior, el Consejo realizó varias recomendaciones a Meta. Cualquier sistema cuya finalidad sea evitar errores debe priorizar la expresión que sea relevante para los derechos humanos, incluida la expresión de importancia pública. Como parte de sus iniciativas por mejorar sus procesos para todos los usuarios, Meta debería tomar medidas para mitigar el daño que provoca el contenido que permanece publicado durante la fase de revisión adicional. Por otra parte, debería mejorar de manera radical la transparencia de sus sistemas.

Conclusiones principales

El Consejo reconoce que el volumen y la complejidad del contenido que se publica en Facebook e Instagram plantea desafíos a la hora de diseñar sistemas que respeten los compromisos de Meta en materia de derechos humanos. No obstante, el sistema de verificación cruzada actual presenta carencias en aspectos fundamentales que la empresa debe abordar:

Trato desigual de los usuarios. El sistema de verificación cruzada protege más a unos usuarios que a otros. Si se determina que la publicación de un usuario que figura en las listas de verificación cruzada de Meta infringe las normas de la empresa, esta permanece en la plataforma a la espera de una revisión adicional. Después, Meta aplica todas las políticas, incluidas excepciones y disposiciones que dependen del contexto, lo que probablemente aumenta las posibilidades de que la publicación siga visible en la plataforma. Por el contrario, existen muchas menos probabilidades de que el contenido del resto de los usuarios llegue a manos de revisores que puedan aplicar toda la gama de reglas de Meta. Este trato desigual es especialmente preocupante si tenemos en cuenta la falta de transparencia en torno a las listas correspondientes al sistema de verificación cruzada de Meta. Es cierto que existen criterios claros en cuanto a la inclusión de socios comerciales y líderes gubernamentales en las listas; sin embargo, el acceso al programa es menos claro para aquellos usuarios cuyo contenido puede resultar importante desde el punto de vista de los derechos humanos (como en el caso de los periodistas y las organizaciones de sociedad civil).

Retraso en la eliminación de contenido infractor. El contenido de los usuarios incluidos en las listas de verificación cruzada que se considera que incumple las reglas de Meta permanece publicado en la plataforma mientras se somete a una revisión adicional. Meta informó al Consejo que, en

promedio, se puede tardar más de cinco días en llegar a una decisión con respecto al contenido que publican estos usuarios. Dicho de otro modo, el sistema de verificación cruzada permite que el contenido que se ha identificado que infringe las reglas de la empresa siga estando disponible en Facebook e Instagram en un momento de máxima viralidad y en el que su potencial para provocar daños es elevado. Es posible que Meta no pueda hacer frente al volumen de contenido seleccionado para la verificación cruzada, por lo que el programa sufre de una acumulación de casos pendientes que retrasa la toma de decisiones.

Incapacidad para hacer un seguimiento de métricas clave. Las métricas que Meta emplea actualmente para medir la efectividad de la verificación cruzada no abarcan todos los aspectos fundamentales. Por ejemplo, Meta no facilitó al Consejo información que demostrara que realiza un seguimiento de si las decisiones que se toman en el contexto de la verificación cruzada son más o menos adecuadas que las que se toman mediante los mecanismos de control de calidad habituales. Sin esa información, es difícil saber si el programa está cumpliendo sus objetivos principales relacionados con tomar decisiones correctas sobre contenido. Tampoco resulta fácil evaluar si la verificación cruzada facilita que Meta se desvíe de sus políticas.

Falta de transparencia en cuanto al funcionamiento de la verificación cruzada. Al Consejo también le preocupa la poca información sobre la verificación cruzada que Meta proporcionó tanto al público como a sus usuarios. Actualmente, Meta no informa a sus usuarios si figuran en listas de verificación cruzada ni comparte públicamente sus procedimientos de creación y auditoría de esas listas. No queda claro, por ejemplo, si las entidades que publican contenido que infringe las normas reiteradamente siguen formando parte de las listas de verificación cruzada debido a su perfil. Esta falta de transparencia impide que el Consejo y el público en general sepan cuáles son todas las consecuencias del programa.

Recomendaciones del Consejo asesor de contenido

A fin de respetar los compromisos de Meta con los derechos humanos y abordar los problemas anteriores, su programa de corrección de errores de mayor impacto en Facebook e Instagram debería ser considerablemente distinto. El Consejo ofrece 32 recomendaciones en este sentido, muchas de las cuales se resumen a continuación.

Dada la voluntad de Meta de mejorar la moderación de contenido para todos los usuarios, debería priorizar la expresión que sea relevante para los derechos humanos, incluida la expresión de importancia pública. Se debería priorizar la inclusión de los usuarios con probabilidades de manifestar este tipo de expresión en las listas de entidades que acceden a revisión adicional por encima de los socios comerciales de Meta. Las publicaciones de estos usuarios deberían revisarse en el marco de un proceso independiente para que no tengan que competir por recursos limitados contra los socios

comerciales de Meta. Si bien el número de seguidores puede ser indicativo del interés del público en la forma de expresión de un usuario, su condición de celebridad o su volumen de seguidores no debería ser el único criterio que se tiene en cuenta a la hora de otorgarle protección adicional. Si los usuarios incluidos en listas de verificación cruzada por su importancia comercial publican contenido infractor con frecuencia, no deberían recibir más protección.

Mejorar radicalmente la transparencia en cuanto al sistema de verificación cruzada y su funcionamiento. Meta debería medir, auditar y publicar métricas clave en relación con su programa de verificación cruzada para que se conozca su efectividad. Sería necesario que la empresa definiera criterios claros públicamente para la inclusión en las listas de verificación cruzada y que los usuarios que cumplan los requisitos puedan solicitar que se los incluya en ellas. Algunas entidades de categorías concretas con acceso a la verificación cruzada (como agentes estatales, candidatos políticos y socios de negocio) deberían incluir una marca pública en su cuenta. Así las personas podrían exigir responsabilidades a los usuarios privilegiados en cuanto a su compromiso de seguir las reglas. Además, dado que aproximadamente un tercio del contenido sometido al sistema de verificación cruzada de Meta no pudo escalarse al Consejo entre mayo y junio de 2022, la empresa debe ofrecer la posibilidad de apelar a este organismo tanto el contenido sometido a la verificación cruzada como aquel que abarcan nuestros documentos rectores.

Reducir los daños provocados por el contenido que permanece en la plataforma para una fase adicional de revisión. El contenido que se considera que infringe las normas de forma grave durante la primera evaluación de Meta debería eliminarse u ocultarse mientras se somete a revisiones adicionales. No debería permitirse que contenido de estas características siga estando visible en la plataforma únicamente porque la persona que lo publicó es un socio comercial o una celebridad. Para garantizar que las decisiones se tomen lo más rápidamente posible, Meta debe dedicar los recursos necesarios para poder asumir la revisión de los volúmenes de contenido que considere que requieren revisión adicional.

II. Solicitud de Meta

1. El Consejo asesor de contenido se enteró de la existencia del sistema de verificación cruzada en 2021 mientras trabajaba en la decisión del caso sobre la suspensión de las cuentas del expresidente de los EE. UU. Donald Trump. Si bien Meta no mencionó dicho sistema en la remisión inicial ni en los materiales que envió al Consejo, describió el programa de verificación cruzada en respuesta a una pregunta del Consejo sobre algún tipo de tratamiento diferencial que pudiera haber recibido la cuenta. Como parte de su decisión correspondiente a mayo de 2021, el Consejo realizó dos recomendaciones que atañen al programa de verificación cruzada:

- "Brindar más información que permita a los usuarios entender y evaluar el proceso y los criterios para la aplicación de la concesión de interés periodístico, incluido cómo se aplica esto a las cuentas influyentes".
- "La empresa también debe explicar con claridad la lógica, los estándares y los procesos implicados en la revisión de verificación cruzada, e informar las tasas de error relativas de las determinaciones que se toman por medio de dicho sistema, en comparación con los procedimientos de aplicación de políticas habituales".

2. En septiembre de 2021, The Wall Street Journal divulgó documentación elaborada por la exempleada y crítica de la empresa Frances Haugen. Según el informe del periódico, el sistema de verificación cruzada eximía a los usuarios más influyentes de Meta de los procesos de moderación de contenido habituales. De acuerdo con The Independent, Frances Haugen afirmó que la empresa le había "mentido" al Consejo respecto del sistema de verificación cruzada "de forma reiterada" durante el caso de Trump. En la documentación interna de Meta que publicó The Wall Street Journal, se observó que algunos de sus empleados consideraban que las prácticas "utilizadas para agregar usuarios a la lista blanca" del sistema de verificación cruzada "no podían justificarse de forma pública". De forma similar, según The Wall Street Journal, los usuarios que podían acceder al sistema de verificación cruzada en ese momento disponían de un intervalo de "autorremediación" de 24 horas para editar o eliminar el contenido infractor y, en consecuencia, evitar cualquier penalización que pudiera imponer Meta.

3. El 21 de septiembre de 2021, después de que The Wall Street Journal publicara los artículos, el Consejo instó a Meta a comprometerse con la transparencia respecto del sistema. Al día siguiente, Meta tuvo una reunión informativa con el Consejo respecto de la verificación cruzada. El Consejo concluyó que "el equipo de Facebook encargado de proporcionar información no había sido completamente explícito en sus respuestas acerca de la verificación cruzada. En algunas ocasiones, Facebook no brindó al Consejo información pertinente, mientras que, en otras instancias, la información que proporcionó estaba incompleta".

4. Poco después de que el Consejo pidiera una mayor transparencia respecto de la verificación cruzada, Meta envió la solicitud de esta opinión de asesoramiento normativo. Luego de resumir brevemente el sistema, Meta describió la verificación cruzada como un programa que "ofrece más niveles de revisión para determinado contenido que nuestros sistemas internos marcan como infractor (por medio de revisión manual o automatizada), con el objetivo de prevenir o minimizar errores de moderación por falsos positivos de mayor riesgo". De acuerdo con Meta, se denomina "falso positivo" a la eliminación equivocada de contenido que no infringe las políticas de contenido que establecen lo que está permitido en Facebook e Instagram.

5. Meta planteó las siguientes tres preguntas al Consejo:

Debido a la complejidad de la moderación de contenido a escala, ¿de qué manera debería Facebook equilibrar el deseo de aplicar las Normas comunitarias de forma justa y objetiva con la necesidad de ser flexibles, tener en cuenta los matices y tomar decisiones específicas de cada contexto dentro del marco de la verificación cruzada?

¿Qué mejoras debería hacer Facebook sobre la manera en que regulamos la revisión secundaria de respuesta temprana de nuestro sistema de verificación cruzada para aplicar de forma justa nuestras Normas comunitarias a la vez que minimizamos la posibilidad de aplicar las políticas excesivamente, conservamos la flexibilidad de la empresa y promovemos la transparencia en el proceso de revisión?

¿Qué criterios debería usar Facebook para determinar a quién incluir en la revisión secundaria de respuesta temprana y a quién debería priorizar nuestro clasificador de verificación cruzada conforme a uno de los muchos factores para garantizar la igualdad de acceso a este sistema y su implementación?

6. El Consejo aceptó la solicitud de Meta el 21 de octubre de 2021. Luego de hacerlo, le envió preguntas a Meta. El Consejo le hizo 74 preguntas a Meta. La empresa respondió 58 de forma completa y 11 de forma parcial, pero no respondió 5. Meta tardó meses en responder algunas de estas preguntas.

7. Además, el Consejo recibió 87 comentarios del público en relación con esta opinión de asesoramiento normativo: nueve de Asia-Pacífico y Oceanía; dos del sur y centro de Asia; doce de Europa; tres de Latinoamérica y el Caribe; tres de Oriente Medio y África septentrional; tres de África subsahariana; y 55 de Canadá y los Estados Unidos. Si quieres leer los comentarios del público que se enviaron en relación con esta opinión de asesoramiento normativo, haz clic aquí. Asimismo, el Consejo llevó a cabo cuatro talleres regionales centrados en el programa de verificación cruzada.

8. Con base en su análisis sobre esta información, una investigación independiente y la participación de partes interesadas, el Consejo responde ahora las preguntas de Meta y proporciona su evaluación del sistema de verificación cruzada. Meta también le dijo al Consejo que realizó cambios significativos en el programa de verificación cruzada el año pasado. El Consejo entiende que estos cambios constituyen, al menos en parte, un intento de dar respuesta a las críticas públicas que el programa recibió. La explicación del Consejo sobre el programa y su análisis de este se basan en el funcionamiento actual del programa que declara Meta. Sin embargo, en ciertas ocasiones, el Consejo hace referencia a su conocimiento sobre prácticas anteriores, ya que estas indican áreas donde probablemente existen riesgos recurrentes.

9. El Consejo exploró si el programa sirve en la práctica para abordar y mitigar consecuencias negativas de acuerdo con las responsabilidades de Meta con los derechos humanos. Este análisis se cimienta en las normas

internacionales sobre derechos humanos y los compromisos y valores declarados de Meta, e involucra preguntas importantes sobre cómo Meta trata a sus usuarios más influyentes y con mayor poder, permite que el contenido circule en sus plataformas y proporciona información al público sobre sus acciones.

III. Sistema de verificación cruzada de Meta

Explicación de Meta sobre los motivos por los que usa el sistema de verificación cruzada

10. Cada día, los usuarios de Facebook e Instagram crean miles de millones de contenidos. Meta constantemente modera contenido, o bien lo examina, evalúa y toma medidas al respecto según sus políticas de contenido. Dichas políticas son las Normas comunitarias de Facebook y las Normas comunitarias de Instagram.

11. De acuerdo con Meta, moderar contenido a esta escala presenta desafíos, y sus revisores y sistemas automatizados a veces eliminan por error contenido que no infringe sus políticas. Meta se refiere a estas decisiones como "falsos positivos". Los falsos negativos son una forma de subaplicación de políticas y se refieren a contenido que infringe las políticas de Meta, pero que se determina que no es infractor en la etapa de revisión. La subaplicación de políticas también incluye los casos de contenido infractor que no se detectan en la revisión manual o automatizada, así como las opciones de diseño de los sistemas que permiten que contenido infractor siga mostrándose después de una primera revisión.

12. El sistema de verificación cruzada solo aborda la sobreaplicación de políticas (o falsos positivos). Por medio de este sistema, Meta retrasa la ejecución de toda medida para aplicar políticas en contenido seleccionado que inicialmente se identificó como infractor para permitir una posible revisión adicional con el fin de evitar falsos positivos.

13. Meta describió la verificación cruzada como una estrategia de prevención de errores que le permite lograr un equilibrio entre evitar que la expresión de los usuarios se marque como falsos positivos y la necesidad de eliminar rápidamente el contenido infractor. Como parte de la solicitud de esta opinión de asesoramiento normativo, Meta destacó la inclusión de "periodistas que informan desde zonas de conflicto y líderes comunitarios que concientizan sobre situaciones de odio o violencia", así como de agentes cívicos cuando "los usuarios tienen un gran interés en conocer lo que opinan sus líderes".

14. Además, el sistema incluye a los usuarios que Meta describe como "socios comerciales". Estos socios cuentan con puntos de contacto exclusivos en Meta. De acuerdo con la empresa, entre dichos usuarios, se encuentran "organizaciones de salud, editores de noticias, personas que se dedican al entretenimiento, músicos, artistas, creadores y organizaciones benéficas". El

Consejo entiende que esta categoría incluye a los usuarios que son propensos a generar dinero para la empresa, ya sea a través de relaciones comerciales formales o porque atraen a usuarios a la plataforma y los mantienen activos allí. El Consejo entiende que "socios comerciales" probablemente también incluya a empresas importantes, partidos y campañas políticos, y celebridades.

15. Meta explicó al Consejo que agrega a los "socios comerciales" al sistema de verificación cruzada para evitar eliminaciones erróneas que limiten la capacidad de los usuarios y anunciantes de llegar a su público y clientes, así como las consecuencias económicas y relativas a la reputación que dichos errores podrían generarle a la empresa. En el caso de estos usuarios, Meta procura "evitar que tanto los socios comerciales de Facebook como la enorme cantidad de usuarios que los siguen sufran experiencias negativas".

16. Meta manifestó que prefiere la subaplicación de políticas a la sobreaplicación de estas en el contenido que atraviesa la etapa de verificación cruzada, ya que "en la situación comercial actual, en general, se considera más importante maximizar los beneficios de dicho sistema (evitar los falsos positivos) que minimizar el costo de dicho programa (es decir, visualizaciones de contenido infractor). El motivo de esto es la percepción de censura". Según la interpretación del Consejo, esto significa que, por razones comerciales, abordar la "percepción de censura" podría priorizarse sobre otras responsabilidades con los derechos humanos pertinentes en el caso de la moderación de contenido.

Cómo funciona el sistema de verificación cruzada

17. Los procesos de moderación de contenido habituales de Meta se aplican a la mayoría de los usuarios. Cuando contenido se identifica como infractor de las políticas de contenido de Meta, la empresa toma una medida para aplicar las políticas correspondientes. Por ejemplo, la eliminación de contenido y la colocación de pantallas de advertencia, según el tipo de infracción. Algunas infracciones también pueden causar penalizaciones en la cuenta, como la suspensión y la eliminación. No obstante, en algunas ocasiones, el contenido recibe un tratamiento diferencial, como en el caso del sistema de verificación cruzada.

18. Meta emplea la expresión "verificación cruzada" para hacer referencia a un programa de prevención de falsos positivos. El programa proporciona más instancias de revisión del contenido antes de que se ejecute alguna medida para aplicar las políticas correspondientes. Durante esta fase adicional de revisión, podrían aplicarse políticas de contenido exclusivas de la etapa de escalamiento, que solo equipos especializados pueden aplicar en Meta. Estas políticas incluyen las concesiones de interés periodístico y basadas en el espíritu de la política, así como todas las reglas que Meta determinó que requieren contexto adicional para aplicarse. Los procesos de revisión de verificación cruzada se inician ante dos conjuntos de circunstancias.

19. Por un lado, la verificación cruzada garantiza una **fase adicional de revisión manual del contenido** publicado por entidades específicas habilitadas siempre que se identifique que este exige la aplicación de políticas conforme a las políticas de contenido de Meta. Meta llama a este proceso **revisión secundaria de respuesta temprana** o **ERSR.** Se denomina "entidad" a cualquiera que use Facebook o Instagram, y pueda publicar contenido, como páginas de Facebook, perfiles de Facebook y cuentas de Instagram. Las entidades pueden representar a personas individuales y a grupos u organizaciones. Meta crea y conserva listas de entidades que, según su decisión, están habilitadas para beneficiarse de la ERSR. Esto significa que, si alguna entidad habilitada publica contenido que se identifica como infractor de las Normas comunitarias, no se eliminará conforme a los procedimientos que se aplican a los usuarios regulares, sino que, en cambio, se someterá a otras instancias de revisión. La ERSR se basa en listas, por lo que solo ciertos usuarios preseleccionados obtienen este beneficio.

20. Por otro lado, el sistema de verificación cruzada ofrece una revisión adicional de cierto contenido que se identifica como infractor de las políticas de Meta, sin importar la identidad del usuario que lo publicó. Meta llama a esto **revisión secundaria general** o **GSR**. Siempre que, mediante revisión manual o automatizada, se identifica como infractor de una política de Meta algún contenido que publica una entidad de la plataforma, la empresa usa un proceso automatizado que se llama "clasificador de verificación cruzada" para analizar al instante varios factores y determinar si el contenido debe someterse a una revisión adicional y cuál debería ser su prioridad en una cola junto con otro contenido que espera el mismo tipo de revisión. De acuerdo con Meta, como este sistema se basa en las características del contenido, todo contenido que publique cualquier usuario de Facebook o Instagram cumple los requisitos para que se lo seleccione para la GSR. La GSR se implementó en 2021, y el Consejo entiende que, en cierta medida, se desarrolló e implementó en la plataforma como respuesta a la crítica hecha a la ERSR, incluidas las revelaciones de Haugen.

21. La detección inicial de contenido en ambos tipos de verificación cruzada que puede dar pie a una revisión puede suceder de forma proactiva, a través de los sistemas automatizados de Meta después de la publicación del contenido, o de forma reactiva, tras reportes de usuarios. Algunas de las medidas para aplicar políticas que pueden dar lugar a una revisión de verificación cruzada son la eliminación de contenido y la colocación de pantallas de advertencia, según el tipo de infracción. Dado que la mayoría de las infracciones contra políticas de contenido pueden causar penalizaciones en la cuenta, como la suspensión o eliminación, estos tipos de aplicación de políticas también se ven afectados. La verificación cruzada se aplica en Facebook e Instagram, salvo en cierto tipo de contenido (p. ej., reels, podcasts) que actualmente no cumple los requisitos del programa. De acuerdo con Meta, "un 10% del contenido orgánico que de cualquier otro modo estaría sujeto a la aplicación

de políticas en relación con la integridad no cumple los requisitos para participar en la revisión de verificación cruzada hoy en día".

22. Luego de que el contenido que cumple los requisitos para la verificación cruzada (a través de GSR o ERSR) se identifica para la aplicación de políticas, pero antes de que quede sujeto al proceso de revisión adicional, el acceso a este contenido sigue siendo total en la plataforma, incluso si en la primera evaluación se determina que el contenido infringe las Normas comunitarias.

23. El Consejo entiende que, si Meta tuviera una mayor disponibilidad de moderadores, más contenido de las colas de revisión de verificación manual recibiría una revisión manual adicional. No obstante, Meta decidió solo garantizar una revisión manual adicional para el contenido que se somete a una ERSR, el sistema correspondiente a las entidades habilitadas. Meta no dedicó los recursos necesarios para que todo el contenido que se identifica a través de la GSR reciba una revisión manual adicional. Si bien los flujos de revisión de estos dos mecanismos son diferentes, como se describe a continuación, si algún revisor en cualquier etapa del proceso determina que el contenido no infringe las políticas de Meta, el proceso de revisión finaliza, y el contenido permanece en la plataforma.

Revisión secundaria de respuesta temprana (ERSR)

24. Meta explicó que, para incluir a las entidades en las listas correspondientes a la ERSR, les asigna una "etiqueta" que se correlaciona con la naturaleza y la sensibilidad de la entidad. Etiquetas específicas corresponden a diferentes listas de la ERSR. Meta precisó que aplica una etiqueta de ERSR a las entidades que pertenecen a las siguientes categorías: (1) entidades cívicas y gubernamentales; (2) acontecimientos mundiales significativos; (3) organizaciones de medios, empresas, comunidades y creadores, incluidos los anunciantes; (4) entidades en las que históricamente se aplicaron políticas de forma excesiva; (5) entidades legales o regulatorias, o entidades en cuyo caso una acción errónea podría presentar riesgos legales para Meta, por ejemplo, en el contexto de un litigio en curso; (6) entidades cuyo contenido está en revisión, es decir, casos en los que la acción de cualquier revisor perjudicaría la deliberación en curso o presentaría un riesgo para Meta. Según Meta, más allá de los factores que usa para determinar si una entidad se ajusta a alguna de las categorías mencionadas anteriormente, como el gasto publicitario o el historial de aplicación de políticas, el derecho a acceder a la ERSR también se determina a través de una evaluación del impacto que un posible error de aplicación de políticas podría tener en la empresa en términos del nivel de liderazgo de la empresa que se vería involucrado en la búsqueda de una solución. En otras palabras, un fundamento clave de la ERSR es evitar provocar a personas que cuentan con los medios para involucrar directamente a ejecutivos sénior o para crear una controversia pública que quizá dichos ejecutivos debieran remediar.

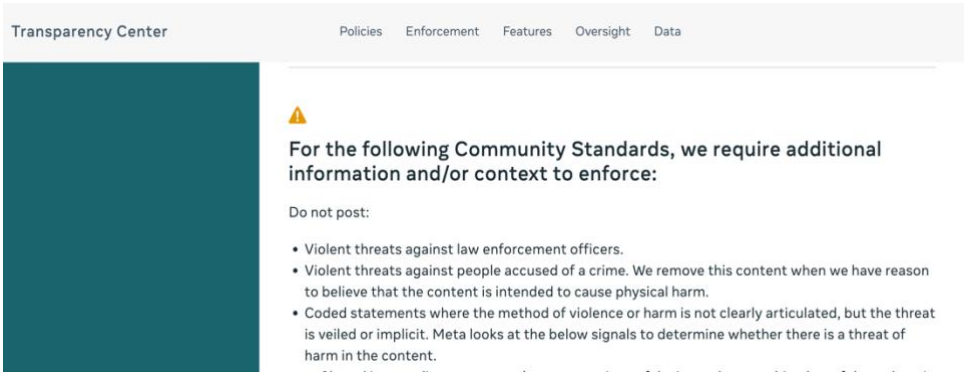
25. Meta le informó al Consejo que, en este momento, está consolidando y actualizando las listas correspondientes a la ERSR. Previamente, las listas de Meta coincidían con el nivel de escalamiento que se necesitaría para aplicar políticas de contenido contra una entidad en particular. De acuerdo con Meta, todas las entidades que actualmente están habilitadas para la ERSR están sujetas al mismo proceso de revisión. Este proceso podría incluir un escalamiento discrecional a los niveles más altos de la empresa.

26. Meta le dijo al Consejo que, durante el segundo trimestre de 2022, estableció criterios generales para agregar entidades a las listas correspondientes a la ERSR y para eliminarlas, así como nuevos procesos para llevar a cabo auditorías y supervisión interna periódicas. Meta no proporcionó detalles sobre estos procesos ni sobre qué acciones podrían causar la reevaluación y remoción de una entidad. Además, explicó que, en general, las etiquetas que colocan a una entidad en una lista de la ERSR expiran en un año, y, teóricamente, las entidades habilitadas deben volver a evaluarse y etiquetarse. De acuerdo con Meta, esta lógica generalmente abarca a las entidades de las siguientes categorías: entidades legales y regulatorias; acontecimientos mundiales significativos; organizaciones de medios; empresas, comunidades y creadores; entidades en las que históricamente se aplicaron políticas de forma excesiva; y entidades cuyo contenido se escala para llevar a cabo una revisión con mayor contexto. Meta mencionó dos excepciones a la regla de expiración de un año. En primer lugar, las etiquetas de las entidades que pertenecen a la categoría "entidades cívicas y gubernamentales" no tienen un período de expiración predeterminado. En segundo lugar, las etiquetas de las entidades de las demás categorías mencionadas anteriormente podrían tener un período de habilitación para la ERSR más breve a discreción de Meta.

27. Cuando, mediante revisión manual o automática, contenido de alguna de las entidades habilitadas se marca para que se apliquen las políticas correspondientes, no se toma ninguna medida al respecto, y en cambio, el contenido se somete a una **fase adicional de revisión por parte de un moderador humano**. Un equipo interno de Meta, al que la empresa denomina "**equipo del mercado regional"**, realiza la primera fase de revisión adicional. Este equipo está conformado por empleados de Meta y contratistas que cuentan con mayor conocimiento contextual y lingüístico sobre un mercado geográfico específico. Si un revisor del equipo del mercado determina que el contenido no es infractor, el proceso finaliza, y el contenido permanece en la plataforma.

28. No obstante, si el revisor del equipo del mercado determina que el contenido infringe las políticas de Meta, el contenido permanece en la plataforma mientras se escala a una nueva instancia de revisión a cargo del "**equipo de respuesta temprana**", según la denominación de Meta. De acuerdo con la empresa, este equipo cuenta con "mayor pericia en las políticas y la capacidad de considerar también contexto adicional".

29. Además, al equipo de respuesta temprana se le permite usar una mayor discreción que a otros moderadores de contenido de Meta, y puede ejecutar políticas de contenido que "requieren más información o contexto para su aplicación". Meta suele marcar estas políticas de contenido con un signo de exclamación en color amarillo en cada norma comunitaria, como se muestra a continuación. Por ejemplo, al final de la norma comunitaria sobre violencia e incitación de Facebook, Meta prohíbe las "amenazas violentas contra personal de las fuerzas del orden". De acuerdo con Meta, la definición de si mantener o eliminar contenido que podría infringir estas partes de la política que dependen del contexto solo puede estar a cargo de un equipo capaz de considerar contexto adicional, como el **"equipo de respuesta temprana"**.



30. El **equipo de respuesta temprana** también podría aplicar lo que Meta llama "concesiones de interés periodístico" y "concesiones basadas en el espíritu de la política", que permiten que contenido que de otro modo sería infractor permanezca en la plataforma porque Meta determina que pertenece al interés público o que, a pesar de que infringe la letra de una política, no incumple la intención de esta. El Consejo también cree que esta discreción alcanza a la aplicación de penalizaciones en la cuenta. Sin embargo, como reveló Meta, el **equipo de respuesta temprana** no cuenta con pericia lingüística ni regional, y se basa en traducciones e información contextual que proporciona el equipo del mercado regional pertinente para evaluar el contenido.

31. En el momento de las reuniones informativas entre el Consejo y Meta, aproximadamente un 0,01% de todo el contenido que, según se identificaba, exigía la aplicación de una política de Meta se escalaba a través del sistema de verificación cruzada a revisores que podían aplicar estas concesiones y políticas contextuales. Se garantiza que el contenido que publican usuarios que están en las listas de la ERSR llegue a esos revisores antes de que se ejecute una medida para aplicar las políticas correspondientes: ni la revisión automática, ni los revisores que trabajan a gran escala ni los revisores del **equipo del mercado** pueden eliminarlo ni colocarle una pantalla de advertencia. Durante todo el período en que el contenido sometido a verificación cruzada espera la definición final, permanece en la plataforma, donde los usuarios pueden indicar que les gusta o compartirlo.

32. Una vez que el **equipo de respuesta temprana** revisa el contenido, si se determina que es infractor, Meta podría implementar medidas para aplicar las políticas correspondientes, como la eliminación del contenido o la colocación de una pantalla de advertencia. Sin embargo, Meta también podría escalar la decisión una instancia más. El Consejo interpreta que los procedimientos de escalamiento en esta etapa son ampliamente discrecionales. Si el equipo de respuesta temprana determina que el contenido "implica la interpretación de un caso extremo de las políticas [de Meta]" o si "presenta un riesgo significativo para la empresa o la comunidad, o si no hay consenso entre las partes internas sobre cómo responder", el equipo de respuesta temprana podría realizar una revisión adicional junto con otros equipos de Meta. De acuerdo con Meta, estas revisiones en etapa de escalamiento "incluyen las perspectivas y los aportes de expertos en el tema de la política de contenido y de los equipos de política pública local, comunicaciones y legal", y podrían incluir contribuciones de otros equipos. Después de dicha revisión, incluso puede escalarse una instancia más al equipo de liderazgo de la empresa antes de que se ejecuten las medidas para aplicar las políticas correspondientes.

33. Asimismo, Meta informó al Consejo que, si "el asunto implica un riesgo significativo de bloqueo de servicios, legal, regulatorio o para la seguridad, o si hay un tiempo limitado para tomar una decisión, en escasas oportunidades, [el equipo de respuesta temprana] escalará la decisión directamente al equipo de liderazgo sénior internacional". Meta indicó que evalúa el riesgo de responsabilidad, la urgencia, el impacto geopolítico, el riesgo de bloqueo de servicios y la falta de consenso entre los equipos internos como factores para escalar estas decisiones.

34. En resumen, el contenido publicado por una entidad que se encuentra en una lista de la ERSR podría someterse hasta a cinco revisiones antes de quedar sujeto a la aplicación de políticas, incluso si revisores reiteradamente determinan que infringe las reglas de Facebook o Instagram, y lo escalan en los canales de los sistemas de verificación cruzada:

1. Revisión inicial, ya sea manual o automatizada, que identifica contenido para la aplicación de políticas conforme a las políticas de Meta.
2. Revisión del equipo del mercado regional.
3. Revisión del equipo de respuesta temprana. Esta es la primera revisión que puede autorizar la aplicación de políticas en el contenido. Este equipo podría solicitar una fase de revisión adicional del equipo de respuesta temprana que incorpore a otros equipos o pasar directamente a una revisión por parte del equipo de liderazgo internacional.
4. Revisión en fase adicional del equipo de respuesta temprana con expertos en la materia, y equipos de política pública, comunicaciones y legal.
5. Revisión del equipo de liderazgo internacional. Se trata de una etapa de escalamiento discrecional del equipo de respuesta temprana sobre la base de la gravedad de las consecuencias para la empresa.

Si en cualquier etapa de revisión se determina que el contenido no es infractor, el proceso se detiene, y el contenido permanece en la plataforma.

35. El canal de la ERSR puede tomar varios días. De acuerdo con Meta, el objetivo interno para que el equipo del mercado complete la revisión de verificación cruzada varía entre 12 y 120 horas, según la gravedad de la posible infracción. En la práctica, Meta indicó que el tiempo medio para tomar la decisión final correspondiente a la revisión secundaria de respuesta temprana es de más de cinco días. En el caso del contenido publicado por usuarios de los Estados Unidos, Meta señaló que le toma "unos 12 días en promedio llegar a una decisión". Otros países tienen tiempos de resolución incluso más lentos. Por ejemplo, el tiempo medio para llegar a una decisión en el caso de Afganistán y Siria es de alrededor de 17 días. En la información que Meta proporcionó al Consejo, el período más prolongado que cierto contenido permaneció en la cola de la revisión secundaria de respuesta temprana fue de 222 días. Meta proporcionó varios gráficos con estos datos al Consejo, donde podía observarse el tiempo medio para llegar a una decisión correspondiente al período entre marzo de 2021 y febrero de 2022, con la tasa de anulaciones y el número de trabajos, o contenidos revisados, y los diferentes países.

36. Meta indicó que la gravedad del contenido que pasa más tiempo a la espera de revisión se designó en algún punto como baja conforme a su "marco de gravedad de las infracciones". Este esquema califica al contenido sobre la base de la norma comunitaria específica que infringe según la primera revisión. El marco de Meta califica cada norma comunitaria de acuerdo con el posible daño que podrían ocasionar las infracciones contra cada política, un cálculo que, según la empresa, realizó sobre la base de una investigación propia. Por ejemplo, considera que el lenguaje que incita al odio es más dañino que el spam, por lo que el posible lenguaje que incita al odio se prioriza sobre el spam en la cola de la ERSR.

37. Dicho eso, en septiembre de 2021, The Wall Street Journal informó que la estrella futbolística brasileña Neymar había publicado imágenes íntimas no consensuadas de otra persona en sus cuentas de Facebook e Instagram. De acuerdo con el informe de The Guardian, el video estuvo online durante más de un día, y, "según una revisión interna de las publicaciones de Neymar, el video obtuvo 56 millones de visualizaciones en Facebook e Instagram antes de su eliminación", a pesar de representar una infracción clara contra las políticas de contenido de Meta. Según la empresa, se pudo acceder a este contenido infractor durante un tiempo prolongado debido a una "demora en la revisión del contenido ocasionada por la acumulación de casos pendientes en ese momento".

38. Una métrica central que Meta le indicó al Consejo que usa para justificar el sistema de verificación cruzada y para evaluar su funcionamiento es la "tasa de anulaciones". Se trata del porcentaje de contenido que se determina que no es infractor durante la revisión de verificación cruzada, por lo que se

revierte la decisión inicial y no se aplican las políticas de contenido que autorizan las reglas de Meta. Meta proporcionó varias cifras al Consejo respecto de la tasa de anulaciones correspondiente al contenido de la ERSR. De acuerdo con Meta, durante distintos períodos a lo largo del año pasado, la tasa de anulaciones varió entre el 30% y el 90%. Cuando la tasa de anulaciones es baja, la ERSR mantiene más contenido que, en última instancia, se determina que es infractor en la plataforma durante las diversas instancias de revisión de verificación cruzada. Cuando la tasa de anulaciones es alta, la ERSR evita en mayor medida que se apliquen políticas por error a contenido no infractor.

39. De acuerdo con Meta, "la mayoría de las visualizaciones ocurren cuando el contenido es nuevo, por lo que la velocidad en las decisiones relacionadas con la revisión y la eliminación rápida del contenido son cruciales para evitar daños". Por lo tanto, el contenido infractor sujeto a la ERSR sigue estando disponible en la plataforma durante todo el período en el que es más probable que reciba la vasta mayoría de sus visualizaciones.

Revisión secundaria general (GSR)

40. El segundo mecanismo que Meta explicó que forma parte de su sistema de verificación cruzada es la **revisión secundaria general**. Mientras que la ERSR se aplica a todo el contenido publicado por entidades habilitadas específicas, la GSR podría aplicarse a cualquier contenido publicado en la plataforma, sin importar quién sea el autor, con base en una determinación algorítmica.

41. La GSR es un sistema relativamente nuevo. Cuando la denunciante de Facebook Frances Haugen divulgó información sobre la verificación cruzada en el otoño boreal de 2021, el Consejo entendió que se refería a la versión previa, que se basaba por completo en la entidad que publicaba el contenido. La empresa ahora también incluye la GSR en la verificación cruzada. Meta precisó que, en 2021, la GSR estaba disponible para contenido que publicaban algunos usuarios y que se implementó de forma completa para el contenido publicado por cualquier usuario a principios de 2022.

42. Para definir qué contenido se envía al canal de la GSR antes de tomar medidas para aplicar las políticas correspondientes, Meta usa un algoritmo (es decir, un conjunto de reglas que una computadora debe seguir para completar una tarea específica) llamado clasificador de verificación cruzada. Este algoritmo se basa en las siguientes características: "sensibilidad del tema (qué tan destacado o sensible es el tema), gravedad de la aplicación de políticas (gravedad de la posible medida que se tomará), probabilidad de aparición de falsos positivos, alcance previsto y sensibilidad de la entidad". Por lo tanto, la sensibilidad de la entidad es un factor en ambos sistemas, aunque en la ERSR, es el factor clave y, en la GSR, es un factor más. Meta indicó que consideró incluir otros factores y prevé hacerlo en el futuro.

43. De acuerdo con Meta, el contenido debe satisfacer dos condiciones para cumplir los requisitos correspondientes a la GSR. Primero, mediante revisión manual o automatizada, se debe haber identificado para la aplicación de políticas (es decir, infractor de una norma comunitaria). Segundo, el clasificador de verificación cruzada debe haberlo marcado como de alta prioridad. Si se cumplen ambas condiciones, no se aplican las políticas en el contenido de forma inmediata, sino que se coloca en una cola para someterse a otra revisión manual por parte del **equipo del mercado regional**. Son los mismos equipos de mercado que también realizan la primera fase de revisión adicional del contenido publicado por entidades habilitadas para la ERSR.

44. Los equipos de mercado no pueden revisar todo el contenido al que se le garantiza una revisión conforme a la ERSR y el contenido que se coloca en una cola para una posible revisión conforme a la GSR. Dado que a las entidades habilitadas que se encuentran en las listas correspondientes a la ERSR se les garantiza una revisión, los equipos de mercado primero deben dedicar sus recursos a este contenido. Con los recursos restantes, el equipo del mercado relevante revisa el contenido de la GSR que se identificó mediante el algoritmo. Los equipos de mercado también revisan cierto contenido fuera del programa de verificación cruzada, entre otras tareas que deben priorizar.

45. Por lo tanto, a pesar de que el algoritmo del clasificador de verificación cruzada podría asignarle una priorización alta al contenido de la GSR y determinar que amerita otra revisión por identificarse como probable falso positivo, es posible que el equipo del mercado no tenga recursos para llevar a cabo la revisión. En algunos casos, si no hay recursos de revisión disponibles en el nivel del equipo del mercado, y Meta decidió poner a disposición sus recursos externos, parte del contenido de la GSR podría enviarse para esa revisión adicional a una revisión manual externa. Si un miembro del equipo del mercado revisa el contenido de la GSR, en la mayoría de los casos, esa decisión es definitiva. Si se determina que el contenido es infractor, por lo general, queda sujeto a la aplicación de políticas (p. ej., se elimina o se coloca una pantalla de advertencia). Si se determina que no infringe las normas, permanece en la plataforma. No obstante, si al **equipo de respuesta temprana** le quedan recursos disponibles luego de cumplir sus obligaciones de revisar todo el contenido de la ERSR antes de su potencial eliminación, es posible que dicho equipo revise el contenido de la GSR de alta prioridad que los revisores del equipo del mercado hubieran marcado como infractor antes de que Meta proceda a la aplicación de políticas.

46. De forma similar a como sucede en el proceso de la ERSR, el contenido que originalmente se evalúa como infractor de las Normas comunitarias y se coloca en la cola de la GSR permanece en la plataforma mientras espera la revisión adicional. Sin embargo, a diferencia de lo que ocurre con la ERSR, el contenido de la cola de la GSR pendiente de revisión no permanecerá en la plataforma de modo indefinido. En última instancia, el contenido que no se revisa "agota el tiempo de espera" de la cola de la GSR. Cuando esto sucede,

Meta retorna a la decisión inicial relacionada con la aplicación de políticas sin otra revisión. Es decir, la medida que se hubiera implementado, como la eliminación o la colocación de una pantalla de advertencia, se aplica de forma aplazada sin otra revisión. Si los revisores no llegan a evaluar cierto contenido de la cola de la GSR, este permanecerá en la plataforma entre dos y cuatro días antes de que Meta lo elimine de la cola de revisión e implemente medidas para aplicar las políticas correspondientes. Al mismo tiempo, el clasificador de verificación cruzada identifica contenido más reciente y de mayor prioridad continuamente, suspende la aplicación de políticas en el contenido y lo agrega a la cola de la GSR.

47. La efectividad general de la GSR se ve limitada por la cantidad de revisores que Meta decide asignar a este tipo de revisión en cada uno de sus mercados. La mayoría de la revisión del contenido de la GSR la lleva a cabo un recurso externo, un revisor del equipo del mercado, o agota el tiempo de espera del sistema. Esto quiere decir que la mayor parte del contenido de la GSR nunca llega al **equipo de respuesta temprana** y, en consecuencia, nunca alcanzará un nivel de revisión en el que pueda aplicarse un análisis del contexto, políticas exclusivas de la etapa de escalamiento y las concesiones de las políticas.

48. Meta también calcula la tasa de anulaciones del contenido que se somete a revisión de verificación cruzada a través del canal de la GSR. Meta proporcionó al Consejo diferentes valores para esta métrica durante el año pasado. En el momento de las reuniones informativas entre el Consejo y Meta, en febrero de 2022, la tasa de anulaciones correspondiente a la revisión secundaria general fue de alrededor del 80%. Más tarde, Meta proporcionó nueva información al Consejo e indicó que, entre marzo y mayo de 2022, la tasa de anulaciones fue del 70% aproximadamente. A pesar de que no permanecieron fijas, estas cifras variaron menos con el tiempo. En las revisiones secundarias, se detectó que la mayor parte del contenido de la GSR que inicialmente se había identificado como infractor no incumplía ninguna política de Meta. Como el tiempo de espera en la cola de la GSR se agota para el contenido, es muy probable que Meta esté aplicando políticas en un gran número de casos de falsos positivos identificados por el clasificador de verificación cruzada.

La verificación cruzada y exenciones reportadas en relación con la aplicación de políticas

49. En el informe de The Wall Street Journal, se describía a la verificación cruzada como un sistema para eximir a "los usuarios vip de la aplicación de políticas habitual de la empresa". Meta le reveló al Consejo que efectivamente cuenta con un sistema que bloquea algunas de las medidas para aplicar políticas fuera del sistema de verificación cruzada. Meta se refiere a esta práctica como "correcciones técnicas", y en el informe público, se la describió como "listas de autorizados" y "listas blancas".

50. Las "correcciones técnicas" son excepciones automáticas a la aplicación de políticas de contenido. Esto significa que invalida casi todos los intentos hechos mediante revisión manual o automatizada de implementar una medida para aplicar políticas para un conjunto preseleccionado de infracciones contra políticas de contenido. Todo el contenido identificado para la aplicación de políticas se evalúa automáticamente para definir si tiene validez alguna "corrección técnica".

51. Si el contenido está protegido por una corrección, se eximirá de esa aplicación de políticas específica. Como Meta explicó, una "corrección técnica" tiene validez solo para una entidad específica en el caso de una infracción contra una política en particular y no sirve para impedir la aplicación de políticas en el caso de otras infracciones contra políticas. En el momento de las reuniones informativas entre el Consejo y Meta, la empresa señaló que aplicaba alrededor de mil **correcciones técnicas** por día. Meta no reveló cuántas entidades ni qué tipo de estas se habían beneficiado de una "corrección técnica".

52. Si el contenido no está protegido por ninguna corrección, se evalúa si reúne los requisitos correspondientes a la verificación cruzada. En ese momento, tenían validez los procesos de verificación cruzada habituales de Meta para identificar si el usuario era una entidad habilitada para la ERSR o si el contenido había sido priorizado por el clasificador de verificación cruzada para la cola de la GSR.

53. Meta primero indicó que principalmente aplica "correcciones técnicas" a "dos grupos de infracciones (spam/comportamiento no auténtico y suplantación de identidad)". Luego, confirmó que, desde el 21 de septiembre de 2022, hay cuatro "correcciones técnicas" activas y que esto podría también cambiar con el tiempo.

54. Meta le dijo al Consejo que sigue habiendo una cantidad limitada de "correcciones técnicas" y que reconoce que hoy por hoy es una necesidad para la empresa. De acuerdo con Meta, tales "correcciones ayudan [a Meta] a evitar errores de aplicación de políticas en contenido o entidades que es muy probable que no infrinjan nuestras políticas y a asignar los recursos de revisión manual a los casos que más lo necesitan".

55. Meta reconoció ciertas deficiencias respecto de sus prácticas pasadas relacionadas con las correcciones técnicas. Meta le dijo al Consejo que "la falta de gestión sobre las prácticas en el pasado, […] de forma involuntaria, tuvo como consecuencia que, en algunas entidades, no se aplicaran muchas de las medidas necesarias". Meta señaló que "distintos equipos podían aplicar diferentes correcciones a la misma entidad de forma tal que, cuando se combinaban, la entidad y su contenido terminaban por no recibir una amplia variedad de medidas para aplicar políticas". Agregó que, dado que esta práctica fue "el resultado inadvertido de un sistema descentralizado,

estamos tomando medidas para asegurar que haya una estructura de gestión en torno al uso de las listas correspondientes a la verificación cruzada".

La verificación cruzada en el contexto de las solicitudes gubernamentales para eliminar contenido

56. Cuando los Gobiernos solicitan que Meta elimine contenido, Meta podría eliminarlo porque infringe las políticas de contenido de la empresa. También podría eliminarlo o "geobloquearlo" por motivos legales y limitar de ese modo su acceso en ciertas áreas. Meta le dijo al Consejo que agrega entidades a las listas correspondientes a la ERSR del sistema de verificación cruzada para protegerlas de medidas erróneas que podrían presentar un riesgo legal para Meta, por ejemplo, en el contexto de un litigio en curso.

57. De acuerdo con Meta, equipos especializados que pueden aplicar las políticas al contenido de inmediato abordan las solicitudes gubernamentales para eliminar contenido, sin importar si la publicación la realizó una entidad habilitada para la ERSR o si el clasificador de verificación cruzada pudiera haberle otorgado una prioridad alta. En otras palabras, las eliminaciones que surgen a partir de solicitudes gubernamentales preceden a los privilegios de la verificación cruzada.

IV. Marco del análisis del Consejo

Normas internacionales sobre derechos humanos

58. El 16 de marzo de 2021, Meta anunció su Política corporativa de derechos humanos, en la que describe su compromiso con el respeto de los derechos de conformidad con los Principios Rectores sobre las empresas y los derechos humanos de las Naciones Unidas (UNGP). Estos principios, respaldados por el Consejo de Derechos Humanos de la ONU en 2011, establecen un marco voluntario que encuadra las responsabilidades de las empresas en cuanto a estos derechos. Estos derechos incluyen, "como mínimo, […] aquellos expresados en la Carta Internacional de Derechos Humanos" (Principio 12).

59. Como empresa mundial que asumió un compromiso con los UNGP, Meta debe respetar las normas internacionales sobre derechos humanos en los lugares donde opera, así como abordar toda consecuencia negativa sobre dichos derechos (Principio 11). Esto también significa que Meta debe "intentar prevenir o mitigar las consecuencias negativas sobre los derechos humanos que guarden un vínculo directo con sus operaciones, productos o servicios prestados por sus relaciones comerciales, incluso si no contribuyeron a generarlas" (Principio 13).

60. En los UNGP, también se establece que las empresas deben llevar a cabo los procesos de debida diligencia respecto de los derechos humanos para evaluar las consecuencias reales y potenciales, y tomar medidas en función

de los resultados (Principio 17). Para hacer esto con efectividad, las empresas deben controlar los indicadores cualitativos y cuantitativos, e incorporar las perspectivas de las partes afectadas (Principio 20).

61. A través de estos casos, el Consejo evalúa las consecuencias de las decisiones específicas relacionadas con la aplicación de políticas en los derechos humanos. Cuando, en estos casos, queda de manifiesto que Meta está provocando consecuencias negativas o quizá no está tomando medidas para identificarlas, controlarlas y limitarlas de forma más general, el Consejo realiza recomendaciones correctivas apropiadas. En una opinión de asesoramiento normativo, el Consejo se centra directamente en las decisiones normativas de Meta, incluidos los procesos de desarrollo y aplicación de políticas, para evaluar si la empresa honra su compromiso con el respeto de los derechos conforme a los UNGP.

62. Para aplicar esto a la verificación cruzada, el Consejo exploró si el programa sirve en la práctica para abordar y mitigar las consecuencias negativas en los derechos humanos de acuerdo con las responsabilidades de Meta. El Consejo también examinó minuciosamente las métricas que Meta usa para determinar la efectividad del programa, además de qué sugiere eso acerca de los objetivos de la empresa.

63. En su análisis, el Consejo determinó que una amplia variedad de derechos podría verse afectada por el programa de verificación cruzada. La libertad de expresión, que incluye el derecho a buscar y recibir información (Artículo 19 del Pacto Internacional de Derechos Civiles y Políticos; observación general n.º 34, 2011, párrafo 11), podría verse potenciada en la medida en que la verificación cruzada ayuda a limitar la aplicación de políticas contra contenido que no infringe las políticas de la plataforma. En consecuencia, se benefician tanto el usuario que publica el contenido como quienes desean acceder a este.

64. El Consejo también destaca que la verificación cruzada podría, en teoría, servir para asegurar que quienes enfrentan obstáculos concretos para ejercer su derecho a la libertad de expresión pudieran beneficiarse de una instancia adicional de protección que el programa podría ofrecer. Por ejemplo, los reportes masivos dirigidos de contenido no infractor podrían verse inhibidos por un sistema de prevención de errores por falsos positivos.

65. No obstante, es posible que el diseño del sistema limite estos efectos positivos si se prepara principalmente para proteger o priorizar la expresión de personas que ya tienen mucho poder. El Consejo también indica que el programa de verificación cruzada genera preocupaciones en torno a la no discriminación, ya que ciertas entidades obtienen protección adicional.

66. Además, es posible que la protección que el programa de verificación cruzada ofrece al contenido infractor contribuya a un entorno que inhiba la expresión de quienes podrían ser el blanco de tal contenido. Es posible que

el conjunto de contenido infractor que pudiera dejarse en la plataforma durante tiempo adicional afecte gravemente a una variedad de derechos humanos, y las consecuencias podrían variar según la situación de los usuarios implicados. Es probable que las personas y grupos que sufren marginación y discriminación experimenten de forma más acentuada las consecuencias negativas en los derechos humanos.

67. El análisis del Consejo contempla estas normas. Sus recomendaciones normativas también reconocen las limitaciones de la capacidad de Meta para moderar contenido a gran escala. Si, mediante sus sistemas de moderación, Meta manejara el contenido de todos los usuarios con mayor precisión, no habría necesidad de crear programas especiales basados en entidades habilitadas para honrar de mejor modo su respeto por los derechos humanos.

Valores de Meta

68. Las normas internacionales sobre derechos humanos definen parámetros para las políticas y prácticas de Meta. Sin embargo, en el marco de dichas normas, las empresas de medios sociales pueden adoptar distintos enfoques en relación con el respeto a los derechos. Los valores de Meta deben guiar a las decisiones discrecionales de la empresa.

69. Meta indicó que tiene cinco valores que influyen en el desarrollo de la aplicación de sus políticas de contenido en Facebook e Instagram. Estos valores son la "expresión", la "autenticidad", la "privacidad", la "seguridad" y la "dignidad". De acuerdo con Meta, la "expresión" es el valor "fundamental" de la empresa. El Consejo interpreta que la verificación cruzada, y un sistema de prevención de errores por falsos positivos en general, principalmente involucra la "expresión", la "privacidad", la "seguridad" y la "dignidad".

70. Un sistema de prevención de errores por falsos positivos que mantiene en la plataforma el contenido que no infringe las políticas de Meta contribuye a consolidar Facebook e Instagram como lugares óptimos para la expresión. En cambio, en la medida en que el sistema de prevención de errores por falsos positivos mantiene el contenido infractor y dañino en la plataforma y facilita su alcance, podría perjudicar la "expresión", la "seguridad", la "privacidad" y la "dignidad" de otras personas. En cuanto a que el sistema privilegia el discurso de algunos en relación con el de otros al retrasar o disminuir la probabilidad de la aplicación de políticas, este tratamiento desigual compromete el valor de "dignidad" de Meta, que se relaciona con la expectativa de que la empresa trate a todos los usuarios de forma imparcial. La empresa debe asegurar que sus sistemas estén estructurados de forma tal que se tengan en cuenta absolutamente todos sus valores.

V. Evaluación del sistema de verificación cruzada

71. En el momento de las reuniones informativas entre el Consejo y Meta, la empresa realizaba aproximadamente 100 millones de intentos de aplicación

de políticas en contenido por día. Con semejante volumen, aunque Meta pudiera tomar decisiones de moderación de contenido con un 99% de precisión, seguiría cometiendo un millón de errores diarios. Los errores de moderación de contenido de Meta incluyen la sobreaplicación y la subaplicación de políticas, lo que significa que Meta elimina contenido no infractor y omite eliminar contenido infractor.

72. En este sentido, el uso que Meta hace del sistema de verificación cruzada se enfrenta a desafíos mayores al moderar volúmenes de contenido tan grandes. El Consejo admite que, en este contexto desafiante, Meta necesita mecanismos para abordar tanto los falsos positivos como los falsos negativos. Sin embargo, la empresa tiene la responsabilidad de abordar estos grandes problemas de modos que beneficien a todos los usuarios, y no solo a un grupo selecto. Toda decisión relacionada con el retraso de la implementación de medidas para aplicar políticas o la exención de estas para ciertos usuarios o contenido debe alinearse con las responsabilidades de Meta con los derechos humanos y sus valores declarados. La verificación cruzada, como funcionaba anteriormente y como lo hace ahora, no lo consigue.

73. El Consejo destaca que Meta incorporó mejoras en este sistema, antes de enviar esta solicitud al Consejo y durante el período en que el Consejo evaluó la verificación cruzada. Sin embargo, varios aspectos del sistema de verificación cruzada no se alinean con la responsabilidad que Meta tiene de identificar y mitigar las consecuencias negativas en los derechos humanos o de respetar los valores de la empresa. Entre ellos:

- Un amplio alcance para lograr objetivos variados y contradictorios que da lugar a la visualización y viralidad de contenido infractor.
- El acceso desigual a políticas y aplicación de políticas discrecionales.
- El hecho de que la inscripción en el programa podría exceder la capacidad disponible.
- La omisión de un seguimiento de las métricas clave para evaluar el programa y mejorarlo.
- La falta de transparencia y auditabilidad en torno a su funcionamiento.

74. A pesar de la significativa preocupación de interés público acerca del sistema, Meta no abordó con efectividad sus componentes problemáticos. En esta sección, el Consejo destaca varios de esos problemas. En las siguientes secciones, hacemos una serie de recomendaciones a Meta para describir cómo un sistema de prevención de errores podría cumplir con los compromisos de la empresa de mejor modo.

<u>Amplio alcance para lograr objetivos variados y contradictorios que da lugar a la visualización de contenido infractor</u>

75. Meta le dijo al Consejo que la revisión secundaria de respuesta temprana existe para "proteger la expresión [y] para aumentar la transparencia y la confianza de la comunidad". Además, en la solicitud que envió al Consejo, hizo hincapié en la inclusión de periodistas y líderes comunitarios en la verificación cruzada. La empresa resaltó que la verificación cruzada garantiza que se preserve la expresión en numerosas situaciones relevantes:

- "Miembros de comunidades marginadas que vuelven a compartir lenguaje que incita al odio infractor dirigido hacia ellos para concientizar al respecto o condenarlo, lo que se hubiera eliminado por error, por infringir las políticas sobre lenguaje que incita al odio".
- "Periodistas que informan en zonas de conflicto, en las que organizaciones designadas están activas, cuyo contenido se hubiera eliminado por error, por infringir nuestras políticas sobre organizaciones y personas peligrosas".
- "Desnudos relacionados con la salud, como una reconstrucción posterior a una mastectomía o fotografías de lactancia materna, que se hubieran eliminado por error, por infringir nuestras políticas sobre desnudos".

76. En una reunión con el Consejo, cuando se les preguntó sobre las consecuencias negativas que podrían surgir sin la ERSR, los representantes de Meta dijeron que un inconveniente, por ejemplo, es que se podría impedir la comunicación y el flujo de información en una crisis, como un desastre natural o una agitación política. Estos puntos destacados en la justificación declarada de Meta para el sistema contrastan de forma considerable con su funcionamiento.

77. El Consejo comparte la preocupación de Meta acerca de la eliminación equivocada de contenido no infractor publicado por personas que quieren concientizar acerca de infracciones contra los derechos humanos, que trabajan para promover la salud de las mujeres y que informan sobre otros temas de interés público. De hecho, las decisiones del Consejo abordaron dichos errores. Meta identifica estos casos como "errores en la aplicación de políticas" solo después de que el Consejo se los hace notar. Algunos ejemplos son la decisión sobre el *cinturón de abalorios* (2021-012-FB-UA), en la que se eliminó por error la expresión de un artista indígena que hacía frente al odio después de varias decisiones equivocadas tomadas mediante revisión manual; la decisión del Consejo sobre la *mención de los talibanes en los informes de noticias* (2022-005-FB-UA) acerca de la eliminación incorrecta de la publicación de un medio de comunicación en la que se informaba sobre una organización designada; y la decisión sobre los *síntomas de cáncer de mama y desnudos* (2020-004-IG-UA), en la que se eliminó por error y de forma automática una publicación que debería haberse beneficiado de la excepción relacionada con la salud a las políticas de desnudos de adultos de Meta.

78. Si bien Meta se centra en las expresiones en riesgo que publican contenido no infractor cuando describe el programa, también indicó que el programa de verificación cruzada es útil para una de las funciones centrales de la

empresa, ya que ocupa un "papel importante en cuanto a la gestión de las relaciones de Facebook con muchos de los socios comerciales". En la misma línea, el marco de sensibilidad de las etiquetas correspondientes a la verificación cruzada, que sustenta el factor de "sensibilidad de la entidad" para la clasificación de la GSR y las etiquetas de la ERSR, se vincula directamente, entre otros factores, con el grado de reacción negativa interna y de reputación que se prevé si cierto contenido se elimina por error. Por ejemplo, para Meta, el riesgo de "escalamiento a los niveles más altos (director ejecutivo, director de operaciones)" se corresponde con una etiqueta de "gravedad extremadamente alta" de la verificación cruzada. La correlación que existe entre la prioridad máxima dentro de la verificación cruzada y las inquietudes en materia de gestión de las relaciones comerciales sugiere que las consecuencias que Meta quiere evitar están principalmente vinculadas con las empresas y no con los derechos humanos.

79. A fin de evaluar cómo Meta prioriza las entidades dentro de la verificación cruzada, el Consejo, de forma reiterada, le pidió a Meta que le compartiera la lista correspondiente a la revisión secundaria de respuesta temprana para poder analizarla. Meta no le proporcionó al Consejo dicha lista. El Consejo no puede evaluar de forma exhaustiva hasta qué punto la empresa cumple sus responsabilidades con los derechos humanos conforme al programa o el perfil de las entidades a las que se les garantiza una fase adicional de revisión si no conoce cómo se implementa el programa y quién se beneficia de él con precisión. Meta argumentó que proporcionar la lista de los usuarios que están sujetos a la verificación cruzada incumpliría las obligaciones legales de la empresa en relación con la privacidad de los usuarios. Sobre la base de asesoramiento jurídico, el Consejo cree y le recalcó a Meta que estas inquietudes podrían haberse mitigado y que podría haber proporcionado información más completa.

80. Casi cinco meses después de la primera vez que el Consejo solicitó esta información, Meta le proporcionó al Consejo una lista con datos consolidados y limitados sobre cada entidad que se encontraba en la lista correspondiente a la revisión secundaria de respuesta temprana en ese momento. Específicamente, Meta solo reveló el tipo de entidad (p. ej., usuario de Instagram, página de Facebook), el país y el idioma asociados, según lo que había seleccionado la propia entidad, y si la empresa consideraba que la entidad era un "socio" y "tenía carácter cívico". No se proporcionó toda la información para todas las categorías de entidades. Por ejemplo, un cuarto de las entidades de Instagram de la lista no seleccionó un país o un idioma específicos en la configuración de su perfil ni tampoco se consideran agentes cívicos ni socios de Meta.[50] Esto significa que, en el caso de estas entidades, Meta solo divulgó la existencia, pero no las identidades o características, de un grupo de usuarios de Instagram que se beneficiaban de la verificación cruzada.

81. Esta divulgación parcial perjudica la capacidad del Consejo de llevar a cabo las responsabilidades de supervisión que se le encomendaron. La

descripción de Meta de la categoría "cívico", por ejemplo, incluye a agentes estatales, funcionarios electos, "influencers cívicos" y candidatos a cargos públicos, entre otros. De forma similar, la categoría "socios" abarca a organizaciones de noticias, celebridades, artistas y muchos más. Por ejemplo, el Consejo no puede evaluar en qué medida los periodistas, los defensores de los derechos humanos y los disidentes de países específicos reciben la misma protección en cuanto a su expresión que los agentes estatales que están inscriptos en la ERSR conforme a la política del programa.

82. Meta le dijo al Consejo que no cuenta con un sistema integral para evaluar de forma sistemática qué periodistas, defensores de los derechos humanos o figuras de la sociedad civil de una zona específica deberían estar sujetos a la ERSR. La inclusión de tales usuarios en la lista se basa en decisiones descentralizadas que toma personal de Meta que la empresa describe como "expertos internos con mucho conocimiento del mercado". Esto aumenta el riesgo de que existan brechas e inconsistencias significativas en cuanto a quién obtiene los niveles de protección adicionales para la expresión que ofrece la ERSR de la verificación cruzada.

83. Los periodistas que publican contenido desde contextos de conflicto, la oposición política que procura un cargo público, las celebridades que publican una amplia variedad de contenido y los socios comerciales que publican contenido para vender bienes constituyen perfiles con riesgos fundamentalmente diferentes desde la perspectiva de la libertad de expresión y los derechos humanos. Dados los problemas que Meta tiene para moderar contenido a gran escala, dentro de las limitaciones actuales, el contenido generado por usuarios debería estar sujeto a una diferente priorización basada en los derechos. Meta describió un sistema que no incluye estrategias o tácticas para garantizar que las personas y la expresión que más necesitan protección la reciban a corto plazo, con el objetivo último de proporcionar una mejor moderación de contenido para todos.

84. Conforme a la ERSR, si el contenido de cualquier entidad habilitada se identifica como infractor y se marca para recibir una revisión adicional, dicho contenido, sin importar el perfil de riesgo, permanece en la plataforma durante su período de viralidad pico inmediatamente posterior a su publicación. Esto tiene mucha relevancia por dos motivos. En primer lugar, el contenido viral se difunde con rapidez en la plataforma en cuestión y en otras. En segundo lugar, una vez que una entidad de gran alcance publica contenido, es inevitable que los usuarios lo graben y vuelvan a compartir individualmente, incluso si se elimina la publicación original. Esto significa que las cuentas que se benefician del sistema de verificación cruzada de la ERSR podrían subir contenido infractor a sabiendas de que puede tener un gran alcance a pesar de infringir las normas.

85. A pesar de que el Consejo destaca que Meta afirmó que tiene un sistema para priorizar el contenido de la ERSR de gravedad alta, este contenido

permanece en la plataforma hasta que se completan todas las revisiones necesarias, en ciertas ocasiones, durante períodos prolongados. Por ejemplo, en el caso de Neymar, es difícil entender cómo imágenes íntimas no consensuadas que se publicaron en una cuenta con más de 100 millones de seguidores no llegaran al principio de la cola para una revisión rápida y de alto nivel si había un sistema de priorización implementado. Dada la gravedad de la naturaleza de la infracción contra las políticas y el impacto en la víctima, este caso pone de manifiesto la necesidad que tiene Meta de adoptar distintos enfoques para el contenido pendiente de revisión y para acortar los plazos de revisión.

86. El retraso de la aplicación de políticas en el contenido infractor es una fuente significativa de daño en el marco del programa de la verificación cruzada. De acuerdo con la propia investigación de Meta, las visualizaciones de los usuarios de contenido infractor como consecuencia del sistema de verificación cruzada se deben a "anulaciones incorrectas y a la demora en la aplicación de políticas cuando no deben realizarse anulaciones, en cuyo caso el proceso de revisión secundaria atrasa la implementación de medidas". La empresa reconoce que ofrecer protección adicional al contenido de ciertos usuarios privilegiados podría enfrentar a otros usuarios con contenido infractor, como lenguaje que incita al odio o publicaciones intimidantes.

87. El contenido al que se ofrece una ERSR de forma automática es diferente del contenido que se identifica y envía para una GSR. Por un lado, como se mencionó antes, el porcentaje del contenido identificado para la ERSR que en última instancia se determina que es infractor pareciera variar. Durante períodos en los que la tasa de anulación es baja, una falla clave en el sistema es su incapacidad para asegurar la pronta eliminación del contenido infractor.

88. Por otro lado, la mayoría del contenido identificado para la GSR en última instancia suele definirse como no infractor. Para este sistema, la tasa de anulaciones parece revelar que hay más problemas de sobreaplicación de políticas a gran escala y que la revisión secundaria, en la mayoría de los casos, permite que contenido no infractor siga estando disponible. Por lo tanto, el Consejo señala que, en la medida en que la GSR preserva la expresión en mayor medida, su impacto se ve limitado por la restricción de recursos que impone Meta.

89. En resumen, el Consejo interpreta que, mientras que Meta caracteriza la verificación cruzada como un programa para proteger voces vulnerables e importantes, pareciera estar estructurado y calibrado de un modo más directo para atender inquietudes comerciales. Si bien el Consejo entiende que Meta es una empresa y debe poder diseñar políticas que atiendan asuntos comerciales, no se debe caracterizar a estas políticas como medidas para mitigar riesgos relacionados con los derechos humanos si no cumplen dicho objetivo. Asimismo, si las decisiones de diseño comerciales de Meta perjudican los derechos humanos, debería identificar y, luego, evitar, mitigar

o detener dichas consecuencias negativas a través de mejoras en el programa.

Acceso desigual a políticas y aplicación de políticas discrecionales

90. La verificación cruzada se diseñó para someter cierto contenido a decisiones de moderación con más matices, a fin de determinar si una excepción o una política especializada podrían tener validez para rechazar la aplicación de políticas. De acuerdo con Meta, "si el contenido que se sometió a verificación cruzada se escala para una fase adicional de revisión, luego podría quedar sujeto a una decisión basada en […] políticas para contextos específicos". La verificación cruzada le permite al "equipo de respuesta temprana" realizar una revisión manual, en la que, según cree el Consejo, puede conferir excepciones a la aplicación de políticas, tanto en relación con el contenido específico como con las penalizaciones contra la entidad en sí. El contenido que se revisa a través de la ERSR obligatoriamente llega a este equipo antes de su posible eliminación, y el contenido que se revisa a través de la GSR tiene una mayor probabilidad de llegar a este equipo.

91. Meta le dijo al Consejo y al público en reiteradas ocasiones que el mismo conjunto de políticas aplica para todos los usuarios. Tales afirmaciones y las políticas de contenido públicas son engañosas, ya que solo un subconjunto pequeño de contenido llega a revisores que tienen la facultad para aplicar todo el conjunto de políticas.

92. Por lo tanto, el derecho a la revisión secundaria de respuesta temprana le proporciona al usuario una ventaja significativa. Significa que una mayor proporción del contenido que decide publicar tiene más probabilidades de permanecer en la plataforma. En el caso del contenido no infractor, se protege de una eliminación equivocada. En el caso del contenido infractor, se permite que permanezca en la plataforma durante el momento en que recibe más visualizaciones, antes de la posterior eliminación.

93. El Consejo también cree que, aparte de la aplicación de las políticas de contenido con mayor discreción, el contenido que se revisa en la etapa de escalamiento podría verse favorecido a partir de la decisión de no aplicar restricciones a las cuentas como se hubiera hecho conforme a los procedimientos habituales. En general, las infracciones contra las políticas de contenido mantienen una correspondencia con "faltas" contra las cuentas, que a su vez se corresponden con consecuencias específicas. De acuerdo con el Centro de transparencia de Meta, las faltas conllevan períodos cada vez más prolongados durante los que las cuentas no pueden publicar contenido. En el caso de faltas graves o repetidas, Meta inhabilita las cuentas.

94. El Consejo preguntó acerca de la aplicación de políticas de forma discrecional y de las consecuencias de esto. Meta respondió que "no contamos con datos estadísticamente significativos en los que se distinga

entre las penalizaciones aplicadas a entidades habilitadas para la verificación cruzada en comparación con aquellas que no lo están", además de que "no conocemos ni pudimos localizar ninguna investigación o análisis" que aborde estas posibles discrepancias. Dado que la verificación cruzada podría eximir a los usuarios de las consecuencias en la cuenta, al Consejo le preocupa que la empresa haya decidido no hacer un seguimiento ni analizar esta información o no la haya revelado al Consejo.

95. De acuerdo con el informe público de The Guardian, después de que Neymar publicó contenido infractor, "no quedó sujeto al procedimiento habitual de Facebook para alguien que publica fotos de desnudos no autorizados, que hubiera sido la eliminación de su cuenta". Este ejemplo se reveló gracias a las declaraciones de la denunciante, y no hay claridad respecto de cuál podría ser la dimensión de estas prácticas. El Consejo también le pidió a Meta que le confirmara las restricciones en la cuenta que aplicó en este caso. En última instancia, la empresa reveló que la única consecuencia fue la eliminación del contenido y que la penalización habitual hubiera sido la inhabilitación de la cuenta. El Consejo destacó que, más tarde, Meta anunció que firmó un acuerdo económico con Neymar para que el futbolista "hiciera streams de juegos exclusivamente en Facebook Gaming y compartiera contenido de video con sus más de 166 millones de fans de Instagram".

96. El acceso desigual a la fase de escalamiento de las revisiones, así como las excepciones a las políticas, son especialmente preocupantes dada la falta de objetividad o transparencia en los criterios para la inclusión en las listas correspondientes a la revisión secundaria de respuesta temprana. Como se mencionó anteriormente, no hay claridad respecto de cómo Meta asegura que quienes tienen más probabilidades de quedar sujetos a la sobreaplicación de políticas o de enfrentar desafíos a la hora de ejercer sus derechos a la libertad de expresión reciban esta protección adicional. Al Consejo le preocupa que aquellos que suelen presentar mayor riesgo, incluidos periodistas y defensores de los derechos humanos (quienes es posible que informen sobre organizaciones peligrosas o documenten abusos gráficos), son quienes tienen menos probabilidades de que se los agregue de forma proactiva a dichas listas, dada la inversión que se necesitaría para localizar a esas personas en todo el mundo.

97. Por otro lado, Meta le explicó al Consejo que cuenta con un equipo exclusivo que se encarga de asegurar que todas las entidades que cumplen los requisitos y representan a funcionarios y organizaciones gubernamentales estén inscriptas en la ERSR. Además, los criterios para incluir a "empresas, organizaciones de medios y creadores" parecen ser más claros. De acuerdo con Meta, un criterio, por ejemplo, es la cantidad específica de gastos o ingresos que una entidad genera en la "familia de apps" de Meta, aunque el monto puede variar con el tiempo.

98. Al Consejo también le preocupa que, al administrar el sistema de verificación cruzada, Meta se centra de forma desproporcionada en los mercados más

lucrativos, en lugar de enfocarse en los contextos que representan mayor riesgo para los derechos humanos, incluida la libertad de expresión. En el momento de las reuniones informativas entre el Consejo y Meta, el 42% del contenido que se sometía al canal de la revisión secundaria de respuesta temprana provenía de los Estados Unidos o Canadá. De forma similar, el 20% de todas las entidades de las listas correspondientes a la ERSR en ese momento pertenecían a esos dos países. En cambio, según Meta, solo el 9% de las "personas activas por mes" en Facebook eran de los Estados Unidos y Canadá. De acuerdo con estos datos, los usuarios radicados en los Estados Unidos y Canadá tienen un acceso desproporcionado a canales de revisión especializada a través de la revisión secundaria de respuesta temprana, que garantizan el acceso a todo el conjunto de políticas de Meta, el análisis del contexto y, probablemente, la posibilidad de recibir penalizaciones a la cuenta no estándar en el caso del contenido infractor.

99. Esta disparidad se correlaciona con el hecho de que los "ingresos promedio por persona" en los EE. UU. y Canadá son los más altos del mundo (una tres veces mayores que en Europa y unas doce veces mayores que en Asia-Pacífico). Esta información resalta los incentivos financieros que perfilan el funcionamiento de la ERSR y refuerza las preocupaciones relacionadas con la equidad. Gracias al diseño del sistema de verificación cruzada, los usuarios que se encuentran en mercados lucrativos con mayor riesgo de repercusiones relacionadas con relaciones públicas para Meta disfrutan de un mayor derecho a la protección de su contenido y expresión que el resto de los usuarios que están en otros lugares.

100. Además, en el caso de la GSR, el clasificador de verificación cruzada prioriza el contenido según ciertos factores, como la "sensibilidad del tema", que potencialmente requieren una evaluación automatizada del texto del contenido. Al Consejo le preocupa que Meta no priorice el entrenamiento de sus procesos automatizados en idiomas que se hablan menos y en mercados menos lucrativos. Si se limita la inversión en la moderación en estos idiomas, se reducirá la capacidad de los algoritmos para identificar temas en dicho contenido. Esto sugiere que los usuarios de estos mercados, incluido el sur global, podrían estar en desventaja cuando se evalúa su elegibilidad para la verificación cruzada por medio de la GSR. De forma similar, Meta reveló que a "un grupo de idiomas lo revisan hablantes no nativos mediante nuestras herramientas de traducción y de detección de insultos". Esto refuerza la preocupación del Consejo de que la verificación cruzada no beneficia a todos los usuarios de igual modo, ni siquiera a través de la GSR.

La inscripción en el programa excede la capacidad disponible

101. La elegibilidad para la ERSR y la GSR persistentemente excede la capacidad de revisión manual que Meta asigna al programa de verificación cruzada. Este desfase entre el volumen de contenido que se designa para una fase adicional de revisión a través de estos sistemas y la cantidad inadecuada de

recursos humanos que se asigna a la tarea representa una falla crítica del sistema.

102. Meta dijo al Consejo que "nunca se tuvo la intención de operar con un volumen constante de casos pendientes, pero las limitaciones en la capacidad operativa y los volúmenes cada vez mayores generaron una acumulación de casos en la revisión secundaria de respuesta temprana. Esos casos pendientes consisten en contenido que se evaluó como de gravedad probablemente baja". A pesar de la declaración de Meta de que no tenía la intención de mantener casos pendientes de forma continua, la empresa no asignó suficientes recursos humanos para satisfacer las necesidades de moderación de contenido de estos programas. Además, como se mencionó antes, no todo el contenido sujeto a demoras en la aplicación de políticas es de gravedad baja.

103. La limitación en la capacidad de revisión manual tiene consecuencias diferentes, pero relacionadas para la revisión secundaria de respuesta temprana y la revisión secundaria general. En el caso de la ERSR, la escasez en la capacidad implica que el contenido permanecerá en la plataforma durante el período en el que más probablemente acumule visualizaciones. Como este contenido permanece en la plataforma hasta que se somete a la fase adicional de revisión, el contenido publicado por usuarios de alto perfil inscriptos en la ERSR que infringe las políticas de Meta permanece en la plataforma durante el período en el que alcanza más visualizaciones. Si bien Meta podría intentar revisar el contenido que podría causar más daño primero, no está claro si lo hace de forma sistemática. Esto pone de manifiesto una decisión relacionada con el diseño de seguir proporcionando protección automática a entidades seleccionadas en función de, en gran medida, criterios comerciales.

104. En el caso de la revisión secundaria general, la limitación en la capacidad podría conllevar dos consecuencias. En primer lugar, es posible que el equipo de respuesta temprana ocupe todo su tiempo con contenido de la ERSR, ya que dicho contenido debe someterse a revisión para que pueda ejecutarse cualquier medida para aplicar las políticas correspondientes. Por lo tanto, dicho equipo no suele contar con disponibilidad para revisar el contenido de la GSR, y el contenido de la GSR no llega a este nivel de revisión fundamental, en el que pueden aplicarse políticas que requieren más contexto y discreción. En segundo lugar, la limitación en la capacidad en el nivel de revisión de los equipos de mercado significa que más contenido de la GSR agota el tiempo de espera en la cola antes de su revisión y se elimina de forma predeterminada. La mayoría de este contenido pareciera ser no infractor de manera regular, por lo que una consecuencia clave de la falta de capacidad asignada al contenido de la revisión secundaria general es que Meta elimina una mayor cantidad de contenido que probablemente no infrinja las normas.

105. Estas fallas agravan las disparidades en el tratamiento de los distintos usuarios en la plataforma. Los usuarios privilegiados inscriptos en la ERSR

tienen más posibilidades de que un moderador que puede analizar el contexto y que dispone de una mayor variedad de excepciones a políticas revise su contenido y lo mantenga, así como de beneficiarse de un sistema que garantiza la visualización incluso del contenido infractor durante un tiempo. Por el contrario, el contenido de los usuarios ordinarios, que podría acceder a la GSR, tiene menos oportunidades de hacerlo, tiene más probabilidades de quedar sujeto a la aplicación de políticas sin una revisión del contexto o sin el uso de excepciones, y, a medida que transcurre el tiempo, es más probable que se elimine a pesar de no ser infractor. Este sistema tiene consecuencias graves en los valores de "expresión", "dignidad", "privacidad" y "seguridad" que Meta proclama reivindicar.

Omisión de un seguimiento de las métricas clave para evaluar el programa y mejorarlo

106. El Consejo evaluó las métricas que Meta usa para justificar y evaluar el programa de verificación cruzada. Las métricas que Meta usa en este momento no dan cuenta de todas las preocupaciones clave y, aparentemente, no generaron cambios cuando se identificaron deficiencias. Asimismo, Meta aún no logró supervisar ni definir objetivos en un conjunto suficientemente amplio de métricas que permitan conocer el panorama completo respecto de cómo funciona el programa y poder establecer metas para mejorarlo en consecuencia.

107. Como mencionamos antes, una métrica que Meta calcula es la tasa de anulaciones, es decir, el porcentaje de contenido sujeto a la verificación cruzada y que, en última instancia, se define como no infractor, a pesar de que, mediante revisión automatizada o manual, inicialmente se hubiera marcado como infractor. Según las propias palabras de Meta, "la tasa de anulaciones es la tasa de eficacia del sistema de verificación cruzada". De acuerdo con la información que Meta proporcionó al Consejo, "se espera que el porcentaje [de anulaciones] sea alto, ya que, si ninguna de las decisiones se anulara a través de la revisión cruzada, se estaría revisando el contenido incorrecto mediante este programa".

108. A pesar de que Meta expresó que la tasa de anulaciones debe ser alta, la empresa sigue proporcionando la máxima protección a los usuarios inscriptos en la revisión secundaria de respuesta temprana. Según las cifras que Meta proporcionó al Consejo, dicha tasa varía de forma significativa. Al proporcionarse esta protección a contenido que no tiene una tasa de anulaciones sistemáticamente alta, se sugiere que Meta podría, sobre la base de sus propios objetivos, estar sometiendo a verificación cruzada el contenido incorrecto.

109. La moderación de contenido a gran escala se ve afectada por la sobreaplicación y la subaplicación de políticas. Meta se centra en la prevalencia de contenido infractor como la principal métrica pública para evaluar cuán efectivas son sus iniciativas relacionadas con la moderación para

eliminar contenido dañino. Esto incluye el contenido que atraviesa las distintas etapas del sistema de verificación cruzada. Meta calcula la prevalencia estimando el porcentaje de todas las visualizaciones del contenido en Facebook o Instagram que fueron visualizaciones de contenido infractor. El uso de la prevalencia como su métrica de éxito general podría animar a Meta a automatizar aún más la eliminación de contenido y limitar la aplicación de políticas basada en el contexto para garantizar una baja prevalencia en la plataforma, sin contar con mecanismos adecuados para evitar eliminaciones de contenido equivocadas a gran escala. La tasa de anulaciones constantemente alta en la GSR, por ejemplo, respalda esa interferencia.

110. El Consejo destaca que Meta no le proporcionó información que demostrara que hace un seguimiento de los datos sobre la precisión de las decisiones que se toman a través del sistema de verificación cruzada. Esto significa que, a pesar de que se supone que el programa garantiza que se tomen decisiones de moderación de contenido precisas, aparentemente, Meta no hace un seguimiento de si las decisiones que se toman a través de los canales de la verificación cruzada son más o menos precisas que aquellas que se toman mediante los mecanismos habituales de control de calidad de la etapa de escalamiento. La existencia de datos precisos sería un indicador clave de la posible influencia de las inquietudes en materia de políticas no relacionadas con el contenido en las decisiones de moderación que se toman en la etapa de verificación cruzada. Al medir el éxito solo en función de la tasa de anulaciones, Meta no considera si las decisiones definitivas son las correctas.

111. Además, Meta señaló que los equipos de mercado regionales y el equipo de respuesta temprana son especializados, por lo que cuentan con un conjunto en particular de habilidades, capacitación y acceso a herramientas internas que les permiten tomar decisiones de moderación en el nivel de la verificación cruzada. Sin embargo, como se describió antes, en ciertos puntos de los canales de la ERSR y de la GSR, es posible que revisores contratados tomen las decisiones. Estos revisores no tienen el mismo acceso ni capacitación que los empleados de Meta. Si el objetivo del programa de verificación cruzada es elaborar las decisiones sobre políticas más precisas posibles para las entidades habilitadas y para el contenido importante, medir la precisión de las decisiones que se toman mediante este sistema en general, pero entre los distintos tipos de revisores en particular, debería ser un principio básico para averiguar si el diseño operativo funciona según lo previsto.

112. Asimismo, como uno de los objetivos de la verificación cruzada es proteger el contenido importante que tiene mayor riesgo de afrontar situaciones de sobreaplicación de políticas, Meta debería enfocarse en desarrollar métodos adicionales para identificar dicho contenido. Meta reveló que, si bien está trabajando activamente para conocer y mitigar las situaciones de sobreaplicación y subaplicación de políticas en poblaciones específicas y zonas problemáticas, aún "se debe definir centralmente cuáles son las poblaciones que las padecen. Al estar pendiente una iniciativa de tal índole, no hay disponible un modo efectivo para elaborar una definición anticipada".

Falta de transparencia y auditabilidad en torno al programa y su funcionamiento

113. Por último, al Consejo le preocupa la poca información sobre este programa que Meta proporcionó al público y a sus usuarios. Esta opinión de asesoramiento normativo surgió porque Meta no divulgó al Consejo información clave sobre este programa en el contexto de su deliberación sobre un caso acerca de un usuario prominente sujeto al sistema de verificación cruzada.

114. En este momento, Meta no informa a los usuarios que están sujetos a la ERSR, el mecanismo basado en entidades de la verificación cruzada. Tampoco informa a los usuarios cuando reportan contenido publicado por una entidad sujeta a dicha verificación. Además, la empresa ofrece poca transparencia sobre los complejos procesos de revisión secundaria de los que se beneficia el contenido sujeto a la verificación cruzada.

115. Asimismo, Meta no comparte públicamente sus procedimientos para la creación de la lista correspondiente a la ERSR y su marco de auditorías. El Consejo no sabe, por ejemplo, si las entidades que publican contenido infractor continuamente siguen formando parte de las listas correspondientes a la revisión secundaria de respuesta temprana debido a su perfil. Meta no dio indicios respecto de si el historial de infracciones o la frecuencia de estas son factores que intervienen en la creación o el mantenimiento de las listas correspondientes a la revisión secundaria de respuesta temprana. La falta de transparencia respecto de las auditorías impide que el Consejo y el público en general sepan cuáles son todas las consecuencias del sistema de verificación cruzada.

Conclusiones sobre el sistema de verificación cruzada

116. El Consejo reconoce que un sistema de prevención de errores podría ser una protección útil contra la eliminación incorrecta de contenido importante. Sin embargo, si el sistema de verificación cruzada no aborda dicha expresión y permite que contenido que infringe las normas gravemente permanezca en la plataforma, el programa genera consecuencias negativas en los derechos humanos que Meta no supervisa ni mitiga. Por lo tanto, el Consejo concluye que, actualmente, ni el diseño ni la implementación del sistema de verificación cruzada satisfacen las responsabilidades de Meta con los derechos humanos ni honran los valores de la empresa.

117. En sus decisiones de casos, el Consejo analiza la prueba de tres partes del Artículo 19 del ICCPR y evalúa si las restricciones que se imponen a la expresión satisfacen los requisitos de legalidad, fin legítimo y necesidad y proporcionalidad.

118. La legalidad se refiere a si las reglas se comunican de forma clara y accesible. La existencia, el propósito y la naturaleza del sistema son poco claros, y esto no puede justificarse dadas las consecuencias significativas que la verificación cruzada tiene en el ejercicio de derechos fundamentales. Las políticas de contenido que se presentan como válidas en todo el mundo y que solo pueden aplicarse con contenido adicional en la etapa de escalamiento, incluso mediante la verificación cruzada, son engañosas.

119. El fin legítimo se refiere a si las restricciones se dirigen a objetivos especificados en el Artículo 19, entre ellos, respetar los derechos de los demás y proteger la seguridad nacional, el orden público y la salud pública. Las métricas con las que la empresa mide la efectividad de sus sistemas de aplicación de políticas sugieren que sus motivaciones se centran sustancialmente en razones comerciales.

120. La necesidad y proporcionalidad se refieren a si las restricciones sobre la expresión constituyen la forma menos invasiva de cumplir con el fin legítimo. Aquí, el Consejo reitera su preocupación respecto del acceso desigual a los beneficios de la verificación cruzada. Meta tiene procesos claros para definir si algunos de sus usuarios son entidades habilitadas, como los agentes estatales y los socios comerciales. Sin embargo, el programa no cuenta con criterios claros para otros usuarios que probablemente publiquen contenido muy valioso para los derechos humanos, por lo que beneficia con menor claridad a los demás, por ejemplo, a los miembros de grupos marginados o víctimas de discriminación. Meta tampoco recopila ni supervisa información sobre si este programa genera resultados más precisos en la práctica. Por último, mediante la verificación cruzada, Meta deja el contenido identificado como infractor en sus plataformas de forma predeterminada. Como cuestión de política, Meta desestima lo que se determinó como una respuesta proporcionada a gran escala para cierto contenido, a menudo, solo con base en inquietudes económicas o en materia de relaciones públicas.

121. Para cumplir con las responsabilidades de Meta con los derechos humanos y honrar los valores de la empresa, el sistema para prevenir situaciones de sobreaplicación de políticas debería estructurarse de un modo sustancialmente diferente de como lo está en este momento.

VI. Recomendaciones para la aplicación de políticas

122. En respuesta a las preguntas que planteó Meta, el Consejo aquí proporciona recomendaciones sobre los sistemas de prevención de errores basados en entidades y los sistemas de prevención de errores basados en contenido que se identifica de forma dinámica. Meta tiene la responsabilidad de abordar los desafíos relacionados con la moderación de contenido de modos que beneficien a todos los usuarios, y no solo a un grupo selecto. Sin embargo, dado el enfoque de esta opinión de asesoramiento normativo, el Consejo aquí hace hincapié en los sistemas de prevención de errores de alcance limitado.

Recomendaciones de gestión para los sistemas de prevención de errores basados en entidades

123. Todo sistema que se base en la elegibilidad de la entidad, como la revisión secundaria de respuesta temprana, debe diseñarse minuciosamente, quedar sujeto a vigilancia y someterse a supervisión continua. Esto debe asegurar que se cumplan los propósitos establecidos y se evalúen factores externos y consecuencias inesperadas que pudieran surgir. Un sistema semejante debe proteger a los usuarios que probablemente publiquen contenido con expresiones de especial importancia desde el punto de vista de los derechos humanos.

124. Es fundamental que Meta deje claro cuáles son sus objetivos y ajuste sus sistemas estrictamente para poder cumplirlos. Además, debe evitar proteger las expresiones que infrinjan sus políticas de contenido o sus compromisos con los derechos humanos. Asimismo, dado que ciertos usuarios podrían beneficiarse a partir de protecciones y canales adicionales para la expresión, la empresa debe proporcionar al público información rigurosa sobre estos procesos para que pueda evaluar de forma adecuada la información y las opiniones que ve en la plataforma.

Usuarios que deben incluirse en los sistemas de prevención de errores basados en entidades

125. Meta afirmó que las categorías de inclusión correspondientes al sistema de prevención de errores basado en entidades abarcan "entidades cívicas y gubernamentales", "acontecimientos mundiales significativos", "organizaciones de medios", "entidades en las que históricamente se aplicaron políticas de forma excesiva" y "comunidades marginadas", "empresas", "creadores", "entidades en etapa de escalamiento para revisión" y "entidades legales y regulatorias".

126. Estas categorías amplias requieren una mayor organización y especificación. A la luz de los compromisos de Meta con los derechos humanos y sus valores declarados, si la empresa opta por implementar un sistema de prevención de falsos positivos basado en entidades, hay ciertas categorías de usuarios que *deben* recibir tal protección, usuarios que *podrían* recibir esta protección y usuarios que *no deben* recibir protecciones dados los riesgos que representan para los derechos humanos.

127. En primer lugar, las entidades que *deben* incluirse son aquellas que probablemente produzcan contenido con expresiones importantes desde el punto de vista de los derechos humanos, incluidos asuntos de importancia pública. Esto beneficia tanto a dichos usuarios como a aquellos que desean tener acceso a la información que comparten.

128. Entre estos usuarios, por ejemplo, deben estar las personas cuyo contenido enfrenta un alto riesgo de sufrir situaciones de sobreaplicación de políticas,

periodistas y organizaciones de medios, funcionarios públicos y candidatos, además de otros agentes cívicos, como defensores de los derechos humanos y de comunidades marginadas. En este sentido, el Consejo entiende que un sistema basado en listas es indicio de la concesión de protecciones adicionales a expresiones críticas, y no simplemente de protecciones según la identidad del hablante. El Consejo reconoce que Meta mantiene varias listas de entidades a las que otorga mayor protección, incluido su registro de periodistas y su planilla de "socios de confianza" de la sociedad civil. Estas entidades existentes y verificadas podrían conformar una fuente a partir de la cual la empresa podría crear un sistema objetivo, mundial y basado en normas de derechos humanos, al que puedan acceder todos quienes generen contenido con expresiones que satisfagan los criterios para la inclusión.

129. En segundo lugar, las entidades que *podrían* incluirse pueden basarse en las prioridades de la empresa, por ejemplo, usuarios con importancia comercial y socios comerciales. Aquí se podrían incluir los anunciantes, las empresas con páginas o grupos que tienen riesgo de afrontar situaciones de sobreaplicación de políticas, los usuarios que representan un riesgo especial relacionado con la reputación para la empresa u otros usuarios que mantienen relaciones comerciales con Meta.

130. En tercer lugar, hay entidades que *no deben* incluirse en ningún sistema de prevención de errores basado en entidades que demore todo el proceso de aplicación de políticas. Aquí se incluyen las entidades y los usuarios que reiteradamente crean o comparten contenido que infringe las políticas o Condiciones del servicio de Meta. Con el fin de implementar esta regla, podría aprovecharse el sistema de aplicación de políticas en la cuenta actual de Meta, que se basa en faltas y penalizaciones. Si los usuarios incluidos en listas de verificación cruzada por su importancia comercial publican contenido infractor con frecuencia, no deberían seguir beneficiándose de un sistema que retrasa la aplicación de políticas. Meta tiene la responsabilidad de identificar a dichos usuarios y excluirlos de los sistemas que le proporciona a su contenido infractor más visibilidad. Si bien el número de seguidores podría ser un indicio legítimo del grado de interés del público en la expresión de un usuario, su condición de celebridad o su volumen de seguidores no debería ser el único criterio para que se lo incluya en un sistema de prevención de errores basado en entidades.

131. La inclusión por parte de Meta de todas las entidades en el mismo sistema hace que compitan directamente por los recursos de revisión limitados. Meta debe priorizar adecuadamente la asignación de recursos en los sistemas de prevención de errores que mitigan perjuicios sobre los derechos humanos. En estas circunstancias, Meta debe garantizar que personal capacitado revise el contenido con consecuencias en los derechos humanos o el interés público, con la posibilidad de tener en cuenta contexto adicional, sin importar si el contenido provino de canales basados en entidades o en contenido.

132. El Consejo le recomienda a Meta tomar medidas para usar canales independientes o crear mecanismos de priorización para diferenciar entre los usuarios que deben incluirse por las responsabilidades de Meta con los derechos humanos y los usuarios que se incluyen debido a prioridades comerciales, dado el diferente riesgo que presentan sus perfiles. Por ejemplo, en el caso de las empresas, es más probable que su contenido se identifique como infractor de las reglas sobre spam, ya que podrían publicar rápidamente contenido comercial. Los usuarios con una gran cantidad de seguidores podrían publicar contenido sobre asuntos importantes de interés público, pero también podrían publicar contenido infractor.

Los responsables de la toma de decisiones deben estar calificados y facultados para tomar decisiones que respeten los derechos

133. Para ser coherente con las recomendaciones del Consejo de principio a fin, Meta debe priorizar sus flujos de trabajo de la revisión secundaria para la prevención de errores de acuerdo con el perfil de riesgo y el valor para los derechos humanos.

134. Equipos que cuenten con pericia en el contexto y el idioma deben revisar el contenido publicado por entidades que Meta *debe* incluir sobre la base de inquietudes relacionadas con los derechos humanos. Este canal de revisión, incluidas sus vías de escalamiento, debe depurarse de consideraciones comerciales. Meta debe tomar medidas para garantizar que este equipo no dependa de los equipos de políticas públicas o relaciones gubernamentales, ni de aquellos a cargo de la gestión de las relaciones con alguno de los usuarios implicados.

135. El flujo dedicado a resolver asuntos explícitamente relacionados con las prioridades comerciales de Meta podría abordar, por ejemplo, cuestiones vinculadas con anuncios, aplicación de políticas, reglas sobre spam, comportamientos y limitación de funciones. Un ejemplo de cuestiones vinculadas a comportamientos es una página comercial a la que se la penalice de forma equivocada por subir imágenes a un ritmo mucho más rápido que el de un perfil normal. Ya sea a través de una menor priorización o la separación en un flujo de trabajo diferente, estas revisiones no deben desplazar recursos destinados a la mitigación relacionada con los derechos humanos.

136. El Consejo destaca que el equipo de respuesta temprana, al que se le permite aplicar excepciones a políticas e interpretar el contexto, no exige que sus revisores cuenten con pericia cultural ni lingüística. De acuerdo con Meta, este toma decisiones en función de notas que le proporcionan los equipos de mercados regionales. Meta reconoció que "basarse en traducciones es deficiente". En este contexto, el Consejo insta a Meta a asegurar pericia cultural y lingüística en estos niveles de revisión. Meta debe considerar incorporar empleados con pericia cultural y lingüística de regiones en riesgo a esos equipos, así como desarrollar procedimientos para incluir personal con tal pericia en los procesos de toma de decisiones.

Pautas para crear y administrar las listas correspondientes a los sistemas de prevención de errores basados en entidades

137. Meta debe establecer criterios claros y públicos para la elegibilidad correspondiente a la prevención de errores basada en entidades. Estos criterios deben diferenciar entre usuarios cuya expresión amerita protección adicional desde una perspectiva relacionada con los derechos humanos, incluida la información de interés público, y usuarios incluidos por motivos comerciales. Por ejemplo, actualmente, Meta define una categoría de la verificación cruzada como "organizaciones de medios, empresas, comunidades y creadores". En esta categoría se encuentran las "organizaciones de salud, los editores de noticias, las personas que se dedican al entretenimiento, los músicos, los artistas, los creadores y las organizaciones benéficas". Los criterios así de amplios no son suficientes. Meta también debe desarrollar criterios basados en patrones de comportamiento infractor o no deseado en la plataforma para evitar otorgar protecciones a usuarios dañinos.

138. Meta no debe agregar entidades a los sistemas de prevención de errores antes de que el proceso sea objetivo, transparente y tenga una buena gestión. Todas las entidades propuestas para agregarse a una lista deben conocer esta posibilidad y se les debe dar la opción de rechazarla si lo desean. Se les debe pedir a aquellas que acepten la inclusión que revisen las reglas sobre contenido de Meta y que vuelvan a comprometerse con su cumplimiento. Si bien el Consejo interpreta la verificación cruzada como un sistema que proporciona beneficios a los usuarios incluidos, Meta debe operar sobre la base de principios que contemplen el consentimiento del usuario.

139. Mediante los criterios públicos claros, también se debe proporcionar un marco para que los usuarios que cumplen los requisitos soliciten proactivamente su inclusión en dichas listas. Meta debe establecer un proceso por el cual los usuarios puedan solicitar recibir protecciones para la prevención de errores por sobreaplicación de políticas si reúnen los criterios articulados por la empresa. Los agentes estatales deben reunir los requisitos para que se los agregue o solicitar su inclusión con base en estos criterios y condiciones, sin consideración a ninguna otra preferencia.

140. Además del cumplimiento con los criterios públicos, el proceso para la inclusión, sin importar si un usuario o Meta lo inicia, debe involucrar lo siguiente: (1) una solicitud de revisión de las políticas de contenido de Meta y un compromiso adicional y explícito de cumplirlas; (2) un reconocimiento de las reglas específicas del programa; y (3) un sistema para informar a los usuarios proactivamente acerca de los cambios en las políticas de contenido de Meta con el objetivo de favorecer el reconocimiento y el cumplimiento.

141. Meta en ocasiones trabaja con la sociedad civil a través de su programa de socios de confianza y otras iniciativas de interacción con partes interesadas

con el fin de recopilar información sobre las entidades que deberían tenerse en cuenta para recibir protección. El Consejo recomienda que Meta afiance su interacción con la sociedad civil para la creación de las listas. Los usuarios deben poder nominar a otros que cumplen los criterios públicos, siempre y cuando los nominados puedan rechazar la inclusión. Esto tiene especial urgencia en los países en que la presencia limitada de la empresa no permite identificar a los candidatos para la inclusión de forma independiente.

142. La creación de las listas, y en particular esta interacción, deben estar a cargo de equipos especializados, independientes de los equipos cuyos mandatos podrían representar conflictos de interés, como los equipos de políticas públicas de Meta. Para garantizar que se cumplan los criterios, personal especializado, con la ventaja de contar con aportes locales, debe asegurar la aplicación objetiva de los criterios de inclusión. Los equipos de políticas públicas suelen interactuar con agentes gubernamentales y presionarlos, por lo que se crean incentivos en conflicto inevitables. Si bien podrían nominar candidatos, no deben ser los responsables de la toma de decisiones.

143. Meta le dijo al Consejo que, actualmente, un empleado de la empresa podría decidir por sí mismo agregar entidades a una lista en particular del sistema de verificación cruzada, y no se exige una revisión de esas decisiones. De cara al futuro, la empresa debe implementar un proceso de revisión objetiva y basada en criterios de todas las entidades que recibirán los beneficios adicionales. Al menos dos personas de distintos equipos deben participar para finalizar la inclusión en cualquier protección basada en listas, y los individuos con relaciones personales o comerciales con las entidades nominadas no deben ser responsables de la toma de decisiones.

Pautas para mantener y auditar las listas correspondientes a los sistemas de prevención de errores basados en entidades

144. Además de establecer criterios claros para la entrada en un programa de protección para la prevención de errores, Meta debe definir criterios y procesos claros para la auditoría y la eliminación. Si en algún momento las entidades ya no cumplen los criterios de elegibilidad, deben eliminarse.

145. Meta le dijo al Consejo que la nueva estructura de gestión propuesta incluye reglas para agregar y eliminar a entidades de las listas, reglas de expiración de las etiquetas, procedimientos de auditorías periódicas y una estructura de vigilancia. Además, Meta reveló que, no obstante, había excepciones a algunas de estas reglas, por ejemplo, las entidades "de carácter cívico y gubernamental" no tienen períodos de expiración predeterminados. Meta también compartió que actualmente audita a un subgrupo limitado de entidades en la revisión secundaria de respuesta temprana a medida que avanza hacia una estructura de listas más simple.

146. El Consejo recomienda que Meta exija al menos una revisión anual de todas las entidades incluidas en cualquier sistema de prevención de errores que

proporcione beneficios a tales entidades. Además, debe haber protocolos claros para acortar ese período si se justifica. De forma similar a lo que sugirió para la inclusión inicial en cualquier sistema basado en listas, el Consejo recomienda que al menos dos personas con estructuras de subordinación independientes participen en auditorías internas.

147. Meta también debe asegurar criterios de eliminación claros para todo programa de protección basado en listas. Uno de los criterios debe ser la cantidad de contenido infractor que la entidad publica. Por ejemplo, podrían basarse en una política de "tres faltas", a menos que Meta haya establecido una penalización más severa para las infracciones en cuestión (p. ej., eliminación de la cuenta a raíz de imágenes íntimas no consensuadas). En un sistema semejante, se debe advertir a las entidades, y eliminarlas de la verificación cruzada cuando alcanzan su última falta, sin importar si la infracción amerita la eliminación de la plataforma en su totalidad. Las entidades deben poder apelar la eliminación y volver a aplicar en el futuro.

148. Por último, el Consejo hace hincapié en que, si bien los procedimientos de auditorías internas constituyen un paso en la dirección correcta, este tipo de auditorías sin vigilancia externa no es suficiente. Las auditorías externas, por parte del Consejo u otros terceros (p. ej., investigadores o miembros de la sociedad civil), son necesarias para evaluar si un sistema de prevención de errores mitiga las consecuencias negativas en los derechos humanos. A pesar de que el Consejo admite que las auditorías externas generan inquietudes graves en torno a la privacidad y la seguridad, cree que Meta puede tomar medidas atenuantes para anonimizar y agrupar los datos a fin de abordar dichas inquietudes.

Algunas entidades que reciben protección adicional deben marcarse de forma pública

149. El Consejo reiteradamente le solicitó a Meta que informara a los usuarios y el público sobre sus políticas y prácticas. Todo sistema de prevención de errores basado en entidades debe ofrecer a la totalidad de los usuarios de la plataforma claridad respecto de cómo Meta aplica sus reglas. Actualmente, los usuarios no saben si están inscriptos en la ERSR. Además, no se les informa a los usuarios que ven y reportan contenido publicado por usuarios inscriptos en la ERSR que el contenido podría estar sujeto a procedimientos de revisión especiales.

150. El Consejo recomienda que se marquen públicamente las cuentas correspondientes a algunas categorías de las entidades protegidas por el sistema. Estas categorías incluyen a todos los agentes estatales y candidatos políticos, socios comerciales, entidades de medios y demás figuras públicas que se incluyan debido al beneficio comercial que le representa a la empresa la prevención de falsos positivos. Así las personas podrán exigir a los usuarios privilegiados que den cuenta de si las entidades protegidas honran su compromiso de seguir las reglas y a Meta que se atenga a los parámetros del programa que se anunciaron públicamente.

151. El Consejo identificó varios riesgos en cuanto a la identificación pública de los usuarios inscriptos en un programa de prevención de errores por falsos positivos. En primer lugar, podría haber un riesgo adicional de que usuarios actuaran con malicia e intentaran conseguir el control de cuentas con protecciones especiales, al saber que el contenido infractor permanecerá en la plataforma algún tiempo. En segundo lugar, algunas categorías de usuarios podrían enfrentar acoso u otros ataques si se percibe que mantienen una relación con la empresa o que reciben una protección especial de su parte.

152. Sin embargo, el Consejo cree que estos riesgos pueden mitigarse y que los beneficios superan a estos posibles daños. Para comenzar, Meta debería invertir los recursos necesarios para mejorar la protección de las cuentas de los usuarios sujetos al sistema de prevención de errores. Meta cuenta con experiencia a la hora de proporcionar más instancias de protección a los periodistas y otras categorías de usuarios. Tales procedimientos podrían adaptarse y usarse en cualquier sistema de prevención de errores basado en entidades en el futuro. Si bien el riesgo de que usuarios actúen con malicia es real, no es insalvable en este contexto. A pesar de que la lista correspondiente a la revisión secundaria de respuesta temprana actualmente no es pública, muchos usuarios ya asumen que las cuentas de alto perfil forman parte del programa de verificación cruzada.

153. Meta no debe identificar a los beneficiarios que son defensores de los derechos humanos, las entidades que se incluyen porque históricamente estuvieron sujetas a situaciones de sobreaplicación de políticas ni aquellos inscriptos porque tienen riesgo de sufrir daños, aunque estos deberían poder optar por que se los identifique. Criterios claros para la inclusión y la separación del programa según diferentes objetivos facilitarán este proceso.

154. Por último, cuando los usuarios reportan contenido publicado por una entidad identificada públicamente como beneficiaria de instancias adicionales de revisión, en la aclaración que acompaña el reporte se debe explicitar que se aplicarán procedimientos especiales y se deben explicar los pasos y que posiblemente el tiempo de resolución sea mayor.

Recomendaciones de gestión para los sistemas de prevención de errores basados en contenido

155. Mientras que un sistema basado en entidades debe incluir a los usuarios que probablemente generen contenido con expresiones que ameriten una mayor protección desde el punto de vista de los derechos humanos y a los usuarios que podrían presentar un riesgo particularmente alto de afrontar situaciones de sobreaplicación de políticas, un sistema basado en contenido procura proteger dicho contenido directamente, sin importar quién lo publicó.

Contenido que debe seleccionarse y priorizarse para los sistemas de prevención de errores basados en contenido

156. De acuerdo con Meta, el sistema de respuesta general "clasifica el contenido en función del riesgo de falsos positivos usando ciertos criterios, como la sensibilidad del tema (qué tan destacado o delicado es el tema), la gravedad de la aplicación de políticas (qué tan grave es la potencial medida que se tomará), la probabilidad de un falso positivo, el alcance previsto y la sensibilidad de la entidad (que se basa, en gran medida, en las listas compiladas descritas anteriormente)".

157. Los factores que tienen más peso para el algoritmo de clasificación son la sensibilidad del tema y la sensibilidad de la entidad. Como se mencionó anteriormente, la sensibilidad de la entidad, entre otros factores, está directamente relacionada con el grado de escalamiento interno que podría generar un error. En este sentido, el clasificador de verificación cruzada de Meta también prioriza el contenido que podría causar daños económicos o de reputación, un objetivo que ya se atiende mediante la ERSR. A pesar de que la GSR podría haberse diseñado para responder a algunas de las críticas hechas sobre el anterior sistema de verificación cruzada exclusivamente basado en entidades, esto sugiere que la empresa sigue priorizando la expresión en función del hablante y no de la importancia de la expresión.

158. El Consejo está de acuerdo con que la elegibilidad universal para un sistema de prevención de errores por falsos positivos es una medida satisfactoria. Sin embargo, un sistema semejante debe priorizar la identificación de contenido al que no esté dirigido también un sistema basado en entidades. Debe proporcionar una mayor protección sobre la base de fundamentos relacionados con los derechos humanos. Si bien Meta podría otorgar protección adicional en los casos en que la sobreaplicación de políticas pudiera amenazar sus intereses comerciales, de forma similar a lo que ocurre en los sistemas basados en listas, no debe hacerlo a costa de sus compromisos con los derechos humanos.

159. Un clasificador algorítmico para un sistema de prevención de falsos positivos podría, por ejemplo, priorizar contenido en función de los tipos de decisiones que a los moderadores automatizados o humanos les resulta difícil tomar a gran escala (p. ej., discursos que históricamente sufrieron situaciones de sobreaplicación de políticas o discursos de comunidades marginadas). Conjuntamente, el algoritmo podría priorizar el orden de revisión de este contenido sobre la base de la gravedad de la posible infracción, la probabilidad de que se trate de un falso positivo y la probabilidad de que se viralice.

160. El Consejo recomienda que, para aumentar el impacto de un sistema de prevención de falsos positivos, Meta debe reservar una porción mínima de la capacidad de revisión de los equipos que pueden aplicar todas las políticas de contenido (p. ej., el equipo de respuesta temprana). Además, debe analizar el

contenido que se somete a revisiones adicionales para obtener estadísticas respecto de en qué situaciones los sistemas de Meta están causando los errores de mayor impacto y priorizar los recursos de revisión en consecuencia.

Correcciones técnicas

161. Meta explicó que las "correcciones técnicas" prohíben completamente la aplicación de políticas a una entidad en particular por una infracción específica. Podría haber motivos comerciales para proporcionar tal protección a un conjunto extremadamente selecto de entidades, pero, con cualquier sistema de ese tipo, existe la posibilidad de eximir a entidades que publican contenido infractor de la aplicación de políticas en la moderación de contenido. Si se usa un sistema semejante, debe quedar sujeto al máximo nivel de escrutinio interno y externo. Las "correcciones técnicas" eximen a determinadas entidades de cierta aplicación de políticas y, sin equivocación, se interpretan como una "lista de autorizados" o "lista blanca", más allá de cuán limitado pueda ser su alcance.

162. Todas las recomendaciones respecto de los programas basados en listas, como criterios más claros y estrictos, procesos de revisión entre distintos equipos para otorgar cualquier excepción y procesos de auditoría para mantener las excepciones, tienen validez aquí. Asimismo, se deben prohibir las excepciones para el contenido que Meta clasifica como infracciones de gravedad alta. Meta debe llevar a cabo auditorías periódicas en el caso de todas las medidas para aplicar políticas que se vean bloqueadas por dichas excepciones. Si, como Meta afirma, esto ocurre unas mil veces por día, debe contar con la capacidad adecuada para hacerlo. Esta auditoría, con información sobre el alcance y la precisión del programa, debe incluirse en los informes trimestrales de transparencia de Meta.

163. Por último, la empresa debe buscar de forma proactiva y periódica excepciones imprevistas o accidentales que pudieran persistir de versiones anteriores de este programa. En sus decisiones, el Consejo reiteradamente señaló casos en los que Meta de forma inadvertida no actualizó o ajustó los sistemas, y las consecuencias de tales brechas en la gestión de un sistema de excepciones podrían ser críticas.

Recomendaciones de gestión generales para los sistemas de prevención de errores

164. Más allá de los cambios generales en la gestión respecto de cómo deben establecerse y auditarse los sistemas de prevención de errores basados en listas y en contenido, el Consejo también recomienda que los procedimientos de tales sistemas se centren en la mitigación de daños y queden sujetos a una supervisión continua que permita seguir aprendiendo y mejorando.

Mitigación del daño tras la identificación de contenido infractor

165. Como la empresa misma reconoce, una de las principales causas de daño del sistema de prevención de errores por falsos positivos de Meta surge de la demora en la aplicación de políticas en el contenido infractor durante el período en que es más probable que reciba visualizaciones. Como se mencionó antes, Meta identificó que los factores más importantes responsables de que los usuarios vean contenido infractor de usuarios inscriptos en la verificación cruzada (o contenido en sus plataformas) son "las anulaciones incorrectas y la demora en la aplicación de políticas cuando no deben realizarse anulaciones, en cuyo caso el proceso de revisión secundaria atrasa la implementación de medidas". El Consejo insta a Meta a tomar medidas para mitigar esos daños.

166. En primer lugar, Meta debe tomar medidas para garantizar que las decisiones en las etapas adicionales de revisión se tomen lo más rápido posible. Deben realizarse inversiones y cambios estructurales para ampliar los equipos de revisión con el objetivo de que haya revisores disponibles trabajando en las zonas horarias pertinentes siempre que contenido se marque para someterse a una fase adicional de revisión manual.

167. En segundo lugar, el Consejo recomienda que Meta use métodos adicionales para evitar que, de forma predeterminada, no se tome ninguna medida para aplicar las políticas correspondientes en el contenido sujeto a fases adicionales de revisión, por ejemplo, usando los medios menos intrusivos, como bajar de categoría el contenido, disminuir su viralidad, ocultarlo o eliminarlo de forma temporal. Si se establecen distintos niveles de priorización o canales para el contenido y las entidades de diferente naturaleza, a Meta le resultará más sencillo aplicar medidas variadas en los distintos tipos de contenido.

168. El contenido que se identifique como infractor en la primera evaluación de Meta y que sea de gravedad alta, por ejemplo, de acuerdo con el marco de Meta, debe eliminarse u ocultarse hasta que se haga la revisión, y no se debe permitir que permanezca en la plataforma y siga acumulando visualizaciones por el simple hecho de que su autor es un socio comercial o una celebridad. La diferencia entre las opciones de aplicación de políticas, como la eliminación, el ocultamiento y la disminución de la clasificación, debe basarse en la gravedad de la infracción. El marco de Meta, en teoría, se diseñó para dar cuenta de la probabilidad de daños a corto plazo y si el contenido se identificó como con alta probabilidad de constituir un error de aplicación de políticas. Si el contenido se oculta por estos motivos, se debe proporcionar a los usuarios en su lugar un aviso que indique que hay una revisión pendiente.

169. En tercer lugar, Meta no puede permitir que se acumulen casos en estos programas. Si existe una cola de contenido para revisión que excede la capacidad, es posible que contenido que podría ser infractor permanezca en la plataforma durante un período prolongado. Si llegar a una decisión toma semanas y se demora la aplicación de políticas, hay entidades habilitadas que quedan funcionalmente eximidas de las reglas.

170. Meta debe dedicar los recursos necesarios para poder asumir la revisión de los volúmenes de contenido que considere que requieren instancias adicionales de revisión. Sin embargo, esto no significa que debe ajustar el algoritmo para que se seleccione menos contenido. La consecuencia de que Meta no asigne suficientes recursos a la revisión no debería ser una demora en la aplicación de políticas al contenido ni eliminaciones totales equivocadas llevadas a cabo por sistemas o revisores que trabajan a gran escala. Meta ideó procesos para priorizar la revisión y asegurar que su fuerza laboral tenga un flujo continuo de contenido para revisar. Dado que actualmente en la GSR, la tasa de anulaciones es constantemente alta, el Consejo cree que más contenido se beneficiaría a partir de esta revisión.

171. En cuarto lugar, Meta no debe priorizar de forma automática la revisión secundaria basada en entidades y hacer que una gran parte de la revisión basada en contenido que se selecciona mediante un algoritmo dependa de una capacidad adicional de revisión.

Asegurar la disponibilidad de recursos de apelación

172. Meta le informó al Consejo que no proporciona una instancia de apelación o revisión de forma sistemática en todos los tipos de contenido. Las apelaciones correspondientes a contenido sujeto al programa de verificación cruzada aparentemente se ven afectadas por la misma inconsistencia.

173. El Consejo entiende que proporcionar instancias de apelación para contenido que ya llegó al nivel de análisis más alto dentro de la empresa podría ser innecesario, ya que la apelación replicaría los mismos canales. Sin embargo, al Consejo le preocupa que es posible que cierto contenido no cumpla los requisitos de elegibilidad para una apelación, a pesar de no haber alcanzado esos niveles más altos. El Consejo cree que Meta podría y debería haber ofrecido más claridad respecto de este punto en las reiteradas ocasiones que el Consejo se lo preguntó.

174. Además, al Consejo le preocupa particularmente esta confusión porque se relaciona con la elegibilidad para apelar casos al Consejo, tanto para los usuarios que quieren que se restaure su propio contenido en la verificación cruzada como para reportar el contenido de otros usuarios que se benefician de la verificación cruzada. De hecho, de acuerdo con Meta, "en los meses de mayo y junio de 2022, un promedio del 35% del contenido del sistema de verificación cruzada […] no podía escalarse al Consejo asesor de contenido". Los usuarios incluidos en las listas correspondientes a la revisión secundaria de respuesta temprana se encuentran entre aquellos que tienen el mayor alcance en la plataforma. Esta situación podría privar al Consejo de algunos de los casos de moderación de contenido más críticos de Facebook e Instagram.

175. Como primer paso, Meta debe ofrecer claridad respecto de la elegibilidad para apelar en general, así como garantizar que el contenido que no llega al nivel más alto de revisión pueda apelarse internamente. En segundo lugar, Meta debe garantizar la posibilidad de apelar al Consejo todo el contenido que el Consejo esté facultado para revisar conforme a sus documentos constitutivos, independientemente de si el contenido alcanzó los niveles máximos de revisión dentro de Meta.

Aprender y mejorar

176. Para cumplir con sus responsabilidades con los derechos humanos, Meta debe supervisar las actividades que realiza y que impactan en ellos de forma periódica. Los resultados de estas revisiones deben guiar a Meta para hacer mejoras en sus políticas y prácticas, con el fin de minimizar los perjuicios sobre los derechos humanos. En este caso, Meta tiene una variedad de métricas relacionadas con el programa de verificación cruzada que ya indican qué debería mejorar la empresa. El Consejo cree que Meta también debe proporcionar al público información sobre cómo funciona este sistema, tanto para cumplir sus responsabilidades respecto de la transparencia como para dar cuenta de las mejoras.

177. En primer lugar, Meta ya calcula la tasa de anulaciones correspondiente a su sistema basado en entidades (revisión secundaria de respuesta temprana) y a su sistema basado en contenido (revisión secundaria general). Meta debe usar las tendencias de las tasas de anulaciones para tomar una decisión fundamentada respecto de si, de forma predeterminada, aplicar la medida original en un plazo más breve o qué otra medida aplicar mientras se espera la revisión. Si las tasas de anulaciones constantemente son bajas para subgrupos de infracciones contra políticas en particular o para contenido en idiomas específicos, por ejemplo, Meta debe calibrar de forma continua la rapidez y el grado de intrusión con los que debe aplicar las medidas.

178. En segundo lugar, Meta dijo al Consejo que llevó a cabo ejercicios de análisis post mortem después de que el equipo de "evaluación de riesgos" de Meta identificara áreas de riesgos u ocurriera un hecho que la empresa reconociera como una falla. El Consejo recomienda que se realicen estas y otras revisiones de forma regular de la verificación cruzada, con base en evaluaciones internas del riesgo que examinen el sistema en acción en los puntos clave mencionados en esta opinión de asesoramiento normativo.

179. En tercer lugar, Meta reveló que una de las categorías que usa en la revisión secundaria de respuesta temprana es "entidades en las que históricamente se aplicaron políticas de forma excesiva". Esto significa que la empresa ya identificó entidades para las que Meta reconoce que no puede aplicar sus políticas de forma coherente y efectiva. Además de otorgar a tales entidades acceso a programas de prevención de errores por sobreaplicación de políticas, Meta debe usar estos datos para tomar decisiones fundamentadas

respecto de cómo mejorar sus prácticas relacionadas con la aplicación de políticas a gran escala. Meta debe medir la sobreaplicación de políticas impuesta sobre estas entidades y usar esos datos para identificar otras entidades para las que suceda lo mismo. La disminución de esa métrica debe ser un objetivo explícito y de gran prioridad para la empresa.

180. Meta debe desarrollar y supervisar métricas adicionales para alinear de mejor modo las estrategias de prevención de errores con las normas de derechos humanos. Por ejemplo, Meta debe establecer nuevas métricas para cuantificar el impacto de dejar contenido infractor en la plataforma. En particular, la empresa debe calcular el número de visualizaciones que acumula cierto contenido que en última instancia se elimina mientras espera una revisión debido a mecanismos de prevención de errores. Meta debe especificar un valor inicial para esta métrica y dar información sobre objetivos para reducirlo.

181. Meta reveló que también toma medidas para abordar algunos asuntos relacionados con la subaplicación de las políticas. Entre ellas: "clasificadores para detectar contenido que probablemente infrinja las políticas; reportes de usuarios que identifican contenido potencialmente infractor; procedimientos de revisión manual en los que los equipos revisan contenido potencialmente infractor; operaciones de revisión anticipada para contenido de alto riesgo (HERO), un sistema en el que personas revisan contenido que se prevé que será viral; y apelaciones de quienes reportan contenido, donde estos usuarios pueden apelar la decisión [de Meta]".

182. El Consejo destaca que las iniciativas, a excepción de las relacionadas con las apelaciones y la aplicación de políticas de forma automatizada a gran escala, tienen un alcance limitado. Además, algunas de estas iniciativas compiten por recursos con la verificación cruzada. Por ejemplo, los equipos de mercados realizan la revisión mediante HERO, y estos también deben dedicar recursos a la verificación cruzada. Además, las HERO solo admiten revisiones de contenido que se espera que sea viral. El Consejo está de acuerdo con que el contenido de gran alcance podría causar más daño, pero cree que también deben desarrollarse iniciativas para mejorar la moderación de forma integral. Meta debe continuar invirtiendo en la detección anticipada y en sistemas de advertencia, así como contratando e involucrando personas que cuenten con pericia local y lingüística en sus operaciones de revisión de contenido confiables y seguras, y en sus iniciativas para la creación de las listas correspondientes a los sistemas de prevención de errores.

VII. Recomendaciones sobre la transparencia

183. El Consejo realizó una serie de recomendaciones sobre cómo Meta debería diseñar y gestionar cualquier programa de prevención de errores por falsos positivos. Las responsabilidades de Meta con los derechos humanos también implican que debe ofrecer transparencia al público respecto de estos programas. Los informes de transparencia deben incluir datos integrales para que los usuarios y el público conozcan cómo funciona el programa y cuáles

podrían ser las consecuencias en el discurso público. Además de las métricas descritas, el Consejo recomienda que Meta incluya lo siguiente:

a. Las tasas de anulaciones correspondientes a los sistemas de prevención de errores por falsos positivos, desglosadas según las opciones de diseño y los equipos de aplicación de políticas (p. ej., mercados, respuesta temprana, contratistas, etc.). Por ejemplo, el Consejo recomendó que Meta creara flujos independientes para las distintas categorías de entidades o contenido en función de su expresión o perfil de riesgo. La tasa de anulaciones debe informarse para todo sistema basado en entidades y en contenido, y se deben incluir las categorías de las entidades o el contenido.
b. El número total y el porcentaje de políticas exclusivas de la etapa de escalamiento que se aplicaron gracias a los programas de prevención de errores por falsos positivos, en relación con el total de las decisiones relacionadas con la aplicación de políticas.
c. Promedio y mediana del tiempo transcurrido hasta la decisión definitiva correspondientes al contenido sujeto a los programas de prevención de errores por falsos positivos, desglosados por país e idioma.
d. Datos agrupados en relación con las listas que se usaron en los programas de prevención de errores, incluido el tipo de entidad y la región.
e. Tasa de eliminaciones erróneas (falsos positivos) de todo el contenido revisado, incluido el daño total generado por estos falsos positivos, calculado como el total previsto de visualizaciones del contenido (es decir, sobreaplicación de políticas).
f. Tasa de decisiones de mantenimiento erróneas (falsos negativos) del contenido, incluido el daño total generado por estos falsos negativos, calculado como el total de visualizaciones que el contenido acumuló (es decir, subaplicación de políticas).

184. El Consejo previamente recomendó que Meta divulgara las tasas de error en general, pero también debe "informar las tasas de error relativas de las determinaciones que se toman por medio de la verificación cruzada, en comparación con los procedimientos de aplicación de políticas habituales". El Consejo cree que el énfasis que Meta pone en la prevalencia, si bien es útil en ciertos contextos específicos, no proporciona los incentivos adecuados a la empresa ni las herramientas apropiadas para que el público entienda cómo funciona el ecosistema de moderación de contenido de Meta.

185. Según Meta dijo al Consejo, "actualmente invertimos en la medición de una métrica agrupada de primera línea que nos permita entender los falsos positivos en todo el sistema y estamos trabajando para desarrollar esta métrica que esperamos poder compartir externamente en nuestros informes de transparencia. Esta métrica sería la contraparte de la medición de falsos negativos que hoy en día se informa a través de métricas de prevalencia". Este es un paso en la dirección correcta, y el Consejo insta a Meta a completar este trabajo lo antes posible.

186. Además de las métricas destacadas en las secciones anteriores, que sirven tanto como valores de referencia para las mejoras como para proporcionar información, Meta debe ofrecer más información básica en su Centro de transparencia respecto del funcionamiento de todo sistema de prevención de errores que use e identifique entidades o usuarios para otorgarles protecciones adicionales. El Consejo entiende que existe la posibilidad de que los usuarios actúen con malicia e intenten eludir la aplicación de políticas, por lo que Meta podría optar por resumir algunos puntos de sus prácticas relacionadas con la aplicación de políticas. El nivel actual de transparencia no es adecuado, y el miedo a que usuarios actúen con malicia no lo justifica.

187. De forma más general, el Consejo señala que ofrecer una mayor transparencia a investigadores externos, en particular, acceso a datos, es un componente esencial para la vigilancia de los sistemas de prevención de errores. En las interacciones con partes interesadas que se llevaron a cabo para este análisis, el Consejo observó preocupación respecto de que Meta procura limitar los programas actuales de acceso a datos dirigidos a terceros externos. Dado que los sistemas como la verificación cruzada requieren realizar compensaciones complejas, investigadores independientes podrían proporcionar a Meta estadísticas valiosas sobre el impacto de sus decisiones. El Consejo cree que Meta debe implementar un canal que les permita a los investigadores externos acceder a datos no públicos sobre el sistema de verificación cruzada. Así, podrían interpretar el programa de forma más completa mediante investigaciones de interés público, además de ofrecer sus propias recomendaciones de mejoras. Si bien se deben tomar medidas mitigatorias para proteger la privacidad de los usuarios, Meta podría y debería facilitar una mayor comprensión del funcionamiento de sus plataformas.

VIII. Anexo con recomendaciones y medidas de implementación

El Consejo realizó varias recomendaciones a Meta en su opinión sobre asesoramiento normativo. En este anexo, se agrupan esas recomendaciones con medidas de implementación para supervisar el progreso de Meta. La empresa debe proporcionar información al Consejo sobre el trabajo que realice en torno a la implementación en sus informes trimestrales. Asimismo, Meta debe organizar una reunión semestral con los responsables de alto nivel para informar al Consejo acerca del trabajo que realice para implementar las recomendaciones de la opinión de asesoramiento normativo.

N.º	**Recomendación**	**Medidas de implementación**
Gestión de la prevención de errores basada en entidades		
1	Meta debe dividir, ya sea mediante distintos canales o priorización, todos los programas de prevención de sobreaplicación de políticas basados en listas en dos sistemas independientes: uno	Meta proporciona al Consejo información en la que detalla cómo divide la inclusión y la operación de estas categorías de entidades. Meta divulga detalles sobre estos sistemas en su Centro de transparencia.

	para proteger la expresión conforme a las responsabilidades de Meta con los derechos humanos, y otro para proteger la expresión que Meta considera que es una prioridad comercial y no entra en dicha categoría.	*Aplicación de políticas*
2	Meta debe asegurar que el canal de revisión y la estructura de la toma de decisiones correspondientes al contenido que tiene consecuencias en los derechos humanos o de interés público, incluidas las vías de escalamiento, no atañen consideraciones comerciales. Meta debe tomar medidas para garantizar que el equipo a cargo de este sistema no dependa de los equipos de políticas públicas o relaciones gubernamentales, ni de aquellos a cargo de la gestión de las relaciones con alguno de los usuarios implicados.	Meta proporciona al Consejo información en la que detalla los canales de toma de decisiones y los equipos involucrados en la moderación de contenido con consecuencias en los derechos humanos o de interés público. *Aplicación de políticas*
3	Meta debe mejorar el modo en que su flujo de trabajo dedicado a cumplir sus responsabilidades con los derechos humanos pone a disposición pericia contextual y lingüística en las fases adicionales de revisión, específicamente, en los niveles de toma de decisiones.	Meta proporciona información al Consejo en la que detalla cómo mejoró su proceso actual y aumentó la pericia lingüística y contextual disponible en el momento en que se consideran las decisiones basadas en contexto y las excepciones a las políticas. *Aplicación de políticas*
4	Meta debe establecer criterios claros y públicos para la elegibilidad correspondiente a la prevención de errores basada en listas. Estos criterios deben diferenciar entre los usuarios que ameritan protección adicional desde una perspectiva relacionada con los derechos humanos y aquellos incluidos por motivos comerciales.	Meta publica un informe o una actualización en el Centro de transparencia donde detalla los criterios de elegibilidad en relación con las fases adicionales de revisión basadas en listas, correspondientes a las distintas categorías de usuarios que participarán en el programa. *Transparencia*
5	Meta debe establecer un proceso para que los usuarios puedan solicitar recibir protecciones para la	Meta implementa un sistema de solicitud transparente y al que se puede acceder de forma sencilla y pública para toda protección contra la

	prevención de errores por sobreaplicación de políticas si reúnen los criterios articulados públicamente por la empresa. Los agentes estatales deben reunir los requisitos para que se los agregue o solicitar su inclusión con base en estos criterios y condiciones, sin consideración a ninguna otra preferencia.	sobreaplicación de políticas basada en listas, y detalla qué propósitos tiene el sistema y cómo evalúa las solicitudes la empresa. Meta incluye anualmente el número de entidades que se inscribieron con éxito en la prevención de errores a través de solicitudes, su país y categoría, en el Centro de transparencia. *Aplicación de políticas*
6	Meta debe asegurar que el proceso para la inclusión basada en listas, sin importar quién inició el proceso (la entidad en sí misma o Meta) involucre, como mínimo: (1) un compromiso adicional y explícito por parte del usuario de cumplir las políticas de contenido de Meta; (2) un reconocimiento de las reglas específicas del programa; y (3) un sistema mediante el cual se le compartan los cambios en las políticas de contenido de la plataforma de forma proactiva.	Meta proporciona al Consejo toda la experiencia del usuario para registrarse en un sistema basado en listas, entre otras cosas, cómo los usuarios se comprometen con el cumplimiento de las políticas de contenido y cómo se les notifican los cambios en las políticas. *Aplicación de políticas*
7	Meta debe afianzar su interacción con la sociedad civil para la creación de las listas y la nominación para estas. Los usuarios y las organizaciones de sociedad civil de confianza deben poder nominar a otros que cumplen con los criterios. Esto tiene especial urgencia en los países en que la presencia limitada de la empresa no permite identificar a los candidatos para la inclusión de forma independiente.	Meta proporciona información al Consejo sobre cómo la empresa interactúa con la sociedad civil para determinar la elegibilidad basada en listas. Meta proporciona datos en su Centro de transparencia, desglosados por país, respecto de cómo muchas entidades se agregan a raíz de la interacción con la sociedad civil, en oposición a la selección proactiva que realiza Meta. *Aplicación de políticas*
8	Meta debe emplear equipos especializados, independientes de influjos políticos o económicos, incluso de los equipos de políticas públicas de Meta, para evaluar las entidades que se incluirán en listas. Para garantizar que se cumplan los criterios, personal especializado, con la ventaja de contar con aportes locales,	Meta proporciona al Consejo documentos internos en los que detalla qué equipos están a cargo de la creación de las listas y dónde se encuentran en la organización. *Aplicación de políticas*

	debe asegurar la aplicación objetiva de los criterios de inclusión.	
9	Meta debe exigir que más de un empleado se vea involucrado en el proceso final de agregar nuevas entidades a cualquier lista correspondiente a los sistemas de prevención de errores por falsos positivos. Estas personas deben trabajar en equipos diferentes, pero relacionados.	Meta le proporciona al Consejo información en la que detalla el proceso mediante el cual las nuevas entidades se agregan a las listas, incluido el número de empleados que deben aprobar la inclusión y los equipos a los que pertenecen. *Aplicación de políticas*
10	Meta debe establecer criterios claros para la eliminación. Uno de los criterios debe ser la cantidad de contenido infractor que la entidad publica. La descalificación debe basarse en un sistema de faltas transparente, en el que se advierta a los usuarios que la continuidad en las infracciones podría generar la eliminación del sistema o de las plataformas de Meta. Los usuarios deben tener la posibilidad de apelar dichas faltas a través de procesos imparciales y de fácil acceso.	Meta le proporciona información al Consejo en la que detalla el límite de medidas para aplicar políticas contra las entidades ante el que se revoca su protección conforme al programa basado en listas, incluidas las notificaciones que se envían a los usuarios cuando reciben faltas que afectan su elegibilidad, cuando se los descalifica, y sus opciones de apelación. También debe facilitar al Consejo datos sobre cuántas entidades se eliminan cada año por publicar contenido infractor. *Aplicación de políticas*
11	Meta debe establecer criterios y procesos claros para las auditorías. Si en algún momento las entidades ya no cumplen los criterios, deben eliminarse de inmediato del sistema. Meta debe revisar todas las entidades incluidas en los sistemas de prevención de errores al menos una vez al año. Además, debe haber protocolos claros para acortar ese período si se justifica.	Meta proporciona al Consejo datos sobre la cantidad y el tipo de entidades, y los motivos de su eliminación de las listas de entidades a raíz de auditorías, junto con un cronograma para llevar a cabo auditorías de forma periódica. *Aplicación de políticas*
Transparencia de las listas		
12	Meta debe marcar públicamente las páginas y cuentas de las entidades que reciben protección sobre la base de listas en las siguientes categorías: todos los agentes estatales y candidatos políticos, todos los socios comerciales, todos los agentes	Meta marca todas las entidades de estas categorías como beneficiarias de un programa de prevención de errores basado en entidades y anuncia el cambio en su Centro de transparencia. *Transparencia*

	de medios y todas las demás figuras públicas que se incluyan debido al beneficio comercial que se ofrece a la empresa al evitarse falsos positivos. Otras categorías de usuarios podrían optar por que se los identifique.	
13	Meta debe notificar a los usuarios que reportan contenido publicado por una entidad identificada públicamente como beneficiaria de instancias adicionales de revisión que se aplicarán procedimientos especiales, y se deben explicar los pasos y que posiblemente el tiempo de resolución sea mayor.	Meta le proporciona al Consejo las notificaciones que envía a los usuarios que reportan contenido de usuarios identificados como beneficiarios de fases adicionales de revisión y le confirma la implementación mundial, así como datos probatorios de que estas notificaciones se muestran sistemáticamente a los usuarios. *Aplicación de políticas*
14	Meta debe notificar a todas las entidades que incluye en listas para recibir fases adicionales de revisión, así como proporcionarles la posibilidad de rechazar la inclusión.	Meta le proporciona al consejo (1) las notificaciones que envía a los usuarios donde les informa su inclusión en un programa de fases adicionales de revisión basado en listas y les ofrece la posibilidad de rechazarla; y Meta (2) públicamente informa en su Centro de transparencia la cantidad anual de entidades, por país, que rechazan la inclusión. *Aplicación de políticas*
Fases adicionales de revisión y priorización		
15	Meta debe reservar una porción mínima de la capacidad de revisión de los equipos que pueden aplicar todas las políticas de contenido (p. ej., el equipo de respuesta temprana) para revisar el contenido que se marque mediante sistemas de prevención de errores basados en contenido.	Meta proporciona al Consejo documentación en la que muestra el proceso por el que consideró esta recomendación y los fundamentos para su decisión de implementarla, y publica esta justificación en su Centro de transparencia. *Aplicación de políticas*
16	Meta debe tomar medidas para garantizar que las decisiones en las etapas adicionales de revisión de los sistemas de prevención de errores que demoran la aplicación de políticas se tomen lo más rápido posible. Deben	Meta proporciona al Consejo datos que demuestran una disminución intertrimestral en el tiempo que se demora para tomar una decisión respecto de todo el contenido que se somete a fases adicionales de revisión, desglosado por categoría para la inclusión y país.

	realizarse inversiones y cambios estructurales para ampliar los equipos de revisión con el objetivo de que haya revisores disponibles trabajando en las zonas horarias pertinentes siempre que contenido se marque para someterse a una fase adicional de revisión manual.	*Aplicación de políticas*
17	Meta no debe demorar todas las medidas que se aplican al contenido que se identifica que posiblemente infringe las normas de forma grave y debe analizar la aplicación de intersticiales o eliminaciones mientras esté pendiente alguna fase adicional de revisión. La diferencia entre la eliminación o el ocultamiento, y la disminución de la clasificación debe basarse en una evaluación del daño y podría basarse, por ejemplo, en la política de contenido que posiblemente se infringió. Si el contenido se oculta por estos motivos, se debe proporcionar a los usuarios en su lugar un aviso que indique que hay una revisión pendiente.	Meta actualiza su Centro de transparencia con el nuevo enfoque adoptado para las medidas correspondientes a la aplicación de políticas que se implementan mientras el contenido se somete a fases adicionales de revisión y proporciona al Consejo información en la que detalla las consecuencias que habrá en la aplicación de políticas en función de criterios específicos para el contenido. Meta comparte datos con el Consejo sobre la aplicación de estas medidas y su impacto. *Aplicación de políticas*
Recursos		
18	Meta no puede permitir que se acumulen casos en estos programas. Sin embargo, Meta no debe lograr incrementar la capacidad de revisión relativa de forma artificial al aumentar el límite del clasificador o hacer que el algoritmo seleccione menos contenido.	Meta proporciona al Consejo datos que demuestran una disminución intertrimestral en la cantidad total de contenido acumulado y en la cantidad de días con contenido acumulado correspondientes a las colas de la revisión de verificación cruzada. *Aplicación de políticas*
19	Meta no debe priorizar de forma automática la revisión secundaria basada en entidades y hacer que una gran parte de la revisión basada en contenido que se selecciona mediante un algoritmo dependa de una	Meta proporciona al Consejo documentos internos en los que detalla la distribución del tiempo y el volumen de revisión entre los sistemas basados en entidades y en contenido. *Aplicación de políticas*

	capacidad adicional de revisión.	
20	Meta debe asegurar que equipos que pueden aplicar excepciones y analizar el contexto revisen el contenido que se somete a alguna instancia adicional de revisión dada su importancia desde un punto de vista de los derechos humanos, incluido el contenido de importancia pública.	Meta proporciona al Consejo información en donde muestra el porcentaje de contenido que equipos que pueden aplicar excepciones y analizar el contexto revisan dado que lo publicó una entidad habilitada o porque un algoritmo identificó que ameritaba una instancia adicional de revisión, desglosado por sistema de prevención de errores (p. ej., GSR frente a ERSR). *Aplicación de políticas*
Barreras automáticas para la aplicación de políticas ("correcciones técnicas")		
21	Meta debe establecer criterios claros para la implementación de barreras automáticas para la aplicación de políticas ("correcciones técnicas") y no debe permitir que estas tengan validez para los casos de infracciones de políticas de contenido de alta gravedad. Al menos dos equipos con estructuras de subordinación independientes deben participar en la concesión de las correcciones técnicas para posibilitar un escrutinio entre distintos equipos.	Meta publica el número de entidades que actualmente se benefician de una "corrección técnica" de forma anual e indica de qué políticas de contenido se impide la aplicación. *Aplicación de políticas*
22	Meta debe llevar a cabo auditorías periódicas para asegurarse de que las entidades que se benefician de las barreras automáticas para la aplicación de políticas ("correcciones técnicas") cumplan con todos los criterios para la inclusión. Al menos dos equipos con estructuras de subordinación independientes deben participar en estas auditorías para posibilitar un escrutinio entre distintos equipos.	Meta proporciona información al Consejo sobre los procesos periódicos de auditoría de las listas. *Aplicación de políticas*
23	Meta debe llevar a cabo auditorías periódicas entre varios equipos para buscar de forma proactiva y periódica barreras imprevistas o accidentales a la aplicación de	Meta publica información anualmente sobre las barreras inesperadas que encuentra a la aplicación de políticas y las medidas implementadas para remediar la causa principal. *Aplicación de políticas*

	políticas que pudieran surgir a partir de errores en el sistema.	
Equidad procesal		
24	Meta debe asegurar que todo el contenido que no alcanza el máximo nivel de revisión interna pueda apelarse a la empresa.	Meta publica información acerca del número de decisiones sobre contenido que se toman mediante canales adicionales de revisión que no cumplían los requisitos para una apelación. Estos datos anuales, desglosados por país, deben especificarse de tal forma que den cuenta del porcentaje de contenido que no se apeló porque llegó a la etapa de revisión del equipo de liderazgo internacional. *Aplicación de políticas*
25	Meta debe garantizar la posibilidad de apelar al Consejo todo el contenido que el Consejo esté facultado para revisar conforme a sus documentos constitutivos, independientemente de si el contenido alcanzó los niveles máximos de revisión dentro de Meta.	Meta confirma públicamente que, para todo el contenido amparado por los documentos constitutivos del Consejo, se otorgan identificadores de apelación al Consejo asesor de contenido, con el fin de que se puedan enviar reclamos a este. Además, proporciona documentación para demostrar qué medidas se tomaron para salvar las brechas relacionadas con la disponibilidad de instancias de apelación. Meta crea un canal accesible para que los usuarios obtengan una reparación oportuna cuando no reciban un identificador de apelación al Consejo asesor de contenido. *Aplicación de políticas*
Aprender y mejorar		
26	Meta debe usar los datos que compila para identificar a las "entidades en las que históricamente se aplicaron políticas de forma excesiva" para que puedan tomarse decisiones justificadas respecto de cómo mejorar sus prácticas de aplicación de políticas a gran escala. Meta debe medir la sobreaplicación de políticas impuesta sobre estas entidades y usar esos datos para identificar otras entidades para las que suceda	Meta proporciona datos al público en los que expone los rechazos intertrimestrales en la sobreaplicación de políticas y documentación que demuestra que el análisis del contenido de las "entidades en las que históricamente se aplicaron políticas de forma excesiva" se usa para disminuir la tasa de sobreaplicación de políticas de forma más general. *Aplicación de políticas*

	lo mismo. La disminución de la sobreaplicación de políticas debe ser un objetivo explícito y de gran prioridad para la empresa.	
27	Meta debe usar las tendencias de las tasas de anulaciones para tomar una decisión fundamentada respecto de si, de forma predeterminada, aplicar la medida original en un plazo más breve o qué otra medida aplicar mientras se espera la revisión. Si las tasas de anulaciones constantemente son bajas para subgrupos de infracciones contra políticas en particular o para contenido en idiomas específicos, por ejemplo, Meta debe calibrar de forma continua la rapidez y el grado de intrusión con los que debe aplicar las medidas.	Meta proporciona al Consejo datos en los que detalla las tasas correspondientes al tiempo de permanencia o que transcurre hasta la eliminación del contenido de la cola, desglosados por país, ámbito de la política y otras métricas relevantes, y describe los cambios realizados anualmente. *Aplicación de políticas*
Mejorar la rendición de cuentas del programa		
28	Meta debe llevar a cabo revisiones periódicas de distintos aspectos de su sistema de fases adicionales de revisión, incluido el contenido con el mayor tiempo de resolución y el contenido infractor de perfil más alto que permanece en la plataforma.	Meta publica los resultados de las revisiones realizadas al sistema de verificación cruzada anualmente, incluidos resúmenes de los cambios hechos a raíz de estas revisiones. *Transparencia*
29	Meta debe informar públicamente las métricas que cuantifican los efectos adversos de la demora en la aplicación de políticas a raíz de los sistemas de fases adicionales de revisión, como las visualizaciones que acumula el contenido que se mantiene en la plataforma como resultado de sistemas de prevención de errores, pero que luego se determina que es infractor. En sus informes públicos, Meta debe determinar los valores iniciales de estas	Meta incluye una o más métricas clave que indican las consecuencias negativas de la demora en la aplicación de políticas a la espera de los mecanismos de las fases adicionales de revisión en el Informe de cumplimiento de las Normas comunitarias, junto con objetivos para disminuir estas métricas y el progreso en relación con el cumplimiento de estos objetivos. *Transparencia*

	métricas y exponer objetivos para disminuirlas.	
30	Meta debe publicar informes regulares de transparencia centrados específicamente en las demoras en la aplicación de políticas en los sistemas de prevención de falsos positivos. Los informes deben incluir datos que les permitan a los usuarios y al público comprender cómo funcionan estos programas y cuáles podrían ser sus consecuencias en el discurso público. Como mínimo, el Consejo le recomienda a Meta incluir lo siguiente: a. Las tasas de anulaciones correspondientes a los sistemas de prevención de errores por falsos positivos, desglosadas según distintos factores. Por ejemplo, el Consejo recomendó que Meta creara flujos independientes para las distintas categorías de entidades o contenido en función de su expresión o perfil de riesgo. La tasa de anulaciones debe informarse para todo sistema basado en entidades y en contenido, y se deben incluir las categorías de las entidades o el contenido. b. El número total y el porcentaje de políticas exclusivas de la etapa de escalamiento que se aplicaron gracias a los programas de prevención de errores por falsos positivos, en relación con el total de las decisiones relacionadas con la aplicación de políticas. c. Promedio y mediana del tiempo transcurrido hasta la decisión definitiva correspondientes al contenido sujeto a los programas de	Meta publica informes de transparencia anuales que incluyen estas métricas. *Transparencia*

	prevención de errores por falsos positivos, desglosados por país e idioma. d. Datos agrupados en relación con las listas que se usaron en los programas de prevención de errores, incluido el tipo de entidad y la región. e. Tasa de eliminaciones erróneas (falsos positivos) respecto de todo el contenido revisado, incluido el daño total generado por estos falsos positivos, calculado como el total previsto de visualizaciones del contenido (es decir, sobreaplicación de políticas). f. Tasa de decisiones de mantenimiento erróneas (falsos negativos) del contenido, incluido el daño total generado por este tipo de error, calculado como el total de visualizaciones que el contenido acumuló (es decir, subaplicación de políticas).	
31	Meta debe ofrecer información básica en su Centro de transparencia respecto del funcionamiento de todo sistema de prevención de errores que use y que identifique entidades o usuarios para otorgarles protecciones adicionales.	Se agrega una sección al Centro de transparencia en la que se explica la matriz de los sistemas de prevención de errores (el Consejo entiende que existe la posibilidad de que los usuarios actúen con malicia e intenten eludir la aplicación de políticas, por lo que Meta podría optar por resumir algunos puntos de sus prácticas relacionadas con la aplicación de políticas). *Transparencia*
32	Meta debe implementar un canal que les permita a los investigadores externos acceder a datos no públicos sobre los programas de prevención de errores por falsos positivos. Así, podrían interpretar el programa de forma más completa mediante investigaciones de interés público, además de ofrecer sus	Meta presenta un canal por el que los investigadores externos pueden obtener datos no públicos sobre los programas de prevención de errores por falsos positivos. *Transparencia*

	propias recomendaciones de mejoras. El Consejo entiende que las inquietudes relacionadas con la privacidad de los datos se deben abordar mediante escrutinios estrictos y agrupación de datos.	