Parecer consultivo de políticas sobre programa de verificação cruzada da Meta

I. Resumo Executivo

Em outubro de 2021, após a publicação de reportagens sobre o programa de verificação cruzada da Meta no Wall Street Journal, o Comitê de Supervisão aceitou uma solicitação da empresa para revisar o sistema de verificação cruzada e fazer recomendações sobre como melhorá-lo. Este parecer consultivo sobre políticas é nossa resposta a essa solicitação. Ele analisa a verificação cruzada com base nos compromissos de direitos humanos e valores declarados da Meta, levantando questões importantes sobre como a empresa trata seus usuários mais poderosos. Quando o Comitê começou a estudar esse parecer consultivo sobre políticas, a Meta compartilhou que, na época, realizava cerca de 100 milhões de tentativas de aplicação de políticas sobre o conteúdo dos usuários todos os dias. Nesse volume, mesmo que ela conseguisse tomar decisões sobre conteúdo com 99% de precisão, ainda cometeria um milhão de erros por dia. Nesse sentido, embora uma análise de conteúdo deva tratar todos os usuários de maneira justa, o programa de verificação cruzada responde a desafios mais amplos na moderação de imensos volumes de conteúdo.

De acordo com a Meta, tomar decisões sobre conteúdo nessa escala significa que conteúdos que não violam suas políticas podem ser removidos por engano. O programa de verificação cruzada visa resolver isso fornecendo camadas adicionais de análise humana para certas publicações inicialmente identificadas como violadoras. Quando os usuários nas listas de verificação cruzada da Meta publicam esse conteúdo, ele não é imediatamente removido como seria para a maioria das pessoas, mas fica pendente para análise humana. A empresa refere-se a esse tipo de verificação cruzada como “Análise Secundária de Resposta Inicial” (ERSR, na sigla em inglês). No fim de 2021, a Meta ampliou a verificação cruzada a fim de incluir publicações sinalizadas especificamente para análise posterior com base no conteúdo, e não na identidade de quem publicou. Ela refere-se a esse tipo de verificação cruzada como “Análise Secundária Geral” (GSR, na sigla em inglês).

Em nossa análise, encontramos várias deficiências no programa de verificação cruzada da empresa. Embora a Meta tenha afirmado ao Comitê que a verificação cruzada visa promover os compromissos dela com os direitos humanos, descobrimos que o programa parece mais diretamente estruturado para resolver as preocupações das empresas. O Comitê entende que a Meta é uma empresa, mas, ao fornecer proteção extra a determinados usuários em grande parte segundo os interesses comerciais, a verificação cruzada permite que o conteúdo que de outra forma seria removido rapidamente permaneça ativo por um período maior, podendo causar danos. Também constatamos que a Meta falhou em rastrear dados sobre se a

verificação cruzada resulta em decisões mais precisas e expressamos preocupação com a falta de transparência em torno do programa.

Em resposta, o Comitê fez várias recomendações à Meta. Qualquer sistema de prevenção de erros deve priorizar a expressão, que é importante para os direitos humanos, incluindo a expressão de importância pública. À medida que a Meta avança para melhorar seus processos para todos os usuários, a empresa deve tomar medidas para mitigar os danos causados pelo conteúdo pendente de análise adicional e aumentar radicalmente a transparência em torno dos seus sistemas.

Principais conclusões

O Comitê reconhece que o volume e a complexidade do conteúdo publicado no Facebook e no Instagram representam desafios para a criação de sistemas que respeitem os compromissos da Meta com os direitos humanos. No entanto, em sua forma atual, a verificação cruzada é falha em áreas importantes que a empresa deve abordar:

Tratamento desigual dos usuários. A verificação cruzada concede maior proteção a determinados usuários do que a outros. Se a publicação de um usuário nas listas de verificação cruzada da Meta for identificada como violadora das regras da empresa, ela permanecerá na plataforma aguardando uma análise mais aprofundada. Então, a Meta aplica toda a sua gama de políticas, incluindo exceções e provisões específicas de contexto, à publicação, provavelmente aumentando as chances dela de permanecer na plataforma. Os usuários comuns, por outro lado, têm muito menos probabilidade de ter seu conteúdo alcançado pelos moderadores que podem aplicar toda a gama de regras da Meta. Esse tratamento desigual é preocupante principalmente pela falta de critérios transparentes das listas de verificação cruzada da empresa. Embora existam critérios claros para incluir parceiros de negócios e líderes governamentais, os usuários cujo conteúdo é importante do ponto de vista dos direitos humanos, como jornalistas e organizações da sociedade civil, têm caminhos menos transparentes de acesso ao programa.

Remoção tardia do conteúdo violador. Quando o conteúdo dos usuários nas listas de verificação cruzada da Meta é identificado como violador das regras da empresa e passa por uma análise adicional, ele permanece totalmente acessível na plataforma. A Meta informou ao Comitê que pode levar, em média, mais de cinco dias para chegar a uma decisão sobre o conteúdo dos usuários presentes nas listas de verificação cruzada. Isso significa que, por causa da verificação cruzada, o conteúdo identificado como violador das regras da Meta fica pendente no Facebook e no Instagram quando está em alta e pode causar danos. Como o volume de conteúdo selecionado para verificação cruzada pode ultrapassar a capacidade de análise da empresa, o programa tem operado em sobrecarga, o que atrasa as decisões.

Falha ao rastrear as métricas principais. As métricas que a Meta usa atualmente para medir a eficácia da verificação cruzada não capturam todas as principais questões. Por exemplo, a empresa não forneceu ao Comitê informações que mostrem que ela acompanha se as decisões por meio de verificação cruzada são mais ou menos precisas do que por meio dos mecanismos normais de controle de qualidade. Sem isso, é difícil saber se o programa está cumprindo seus objetivos principais de gerar decisões corretas de moderação de conteúdo ou avaliar se a verificação cruzada resulta em um desvio das políticas da Meta.

Falta de transparência sobre como funciona a verificação cruzada. O Comitê também questiona as informações limitadas que a Meta fornece ao público e aos usuários sobre a verificação cruzada. Atualmente, a Meta não informa aos usuários que eles estão em listas de verificação cruzada e não compartilha publicamente seus procedimentos para criar e auditá-las. Não está claro, por exemplo, se as entidades que publicam continuamente conteúdo violador são mantidas em listas de verificação cruzada com base no perfil. Essa falta de transparência impede que o Comitê e o público entendam todas as consequências do programa.

As recomendações do Comitê de Supervisão

Para cumprir os compromissos de direitos humanos da Meta e enfrentar esses problemas, um programa que corrija os erros de maior impacto no Facebook e no Instagram deve ser estruturado de maneira substancialmente diferente. O Comitê fez 32 recomendações nessa área, muitas das quais estão resumidas abaixo.

Em busca de melhorar sua moderação de conteúdo para todos os usuários, a Meta deve priorizar a manifestação que é importante para os direitos humanos, inclusive a manifestação que é de especial importância pública. Os usuários que provavelmente produzirão esse tipo de manifestação devem ter prioridade na inclusão das listas de entidades que recebem análise adicional, e não os parceiros de negócios da Meta. As publicações desses usuários devem ser analisadas em um fluxo de trabalho separado para que não compitam com os parceiros de negócios da Meta por recursos limitados. Embora o número de seguidores possa indicar utilidade pública na manifestação de um usuário, a quantidade de celebridades ou seguidores de um usuário não deve ser o único critério para receber proteção adicional. Se os usuários incluídos devido à sua importância nos negócios publicarem frequentemente conteúdo violador, eles não devem mais se beneficiar da proteção especial.

Aumentar radicalmente a transparência sobre a verificação cruzada e seu funcionamento. A Meta deve avaliar, auditar e publicar as principais métricas do programa de verificação cruzada para confirmar se ele funciona de forma eficaz. A empresa deve definir critérios claros e públicos para

inclusão nas listas de verificação cruzada, e os usuários que atendem a esses critérios devem poder se inscrever no programa. Algumas categorias de entidades protegidas por verificação cruzada, incluindo agentes estatais, candidatos políticos e parceiros de negócios, também devem ter suas contas marcadas publicamente. Isso permitirá que o público responsabilize os usuários privilegiados se as entidades protegidas mantiverem o compromisso de seguir as regras. Além disso, como cerca de um terço do conteúdo no sistema de verificação cruzada da Meta não pôde ser escalonado ao Comitê de maio a junho de 2022, a empresa deve garantir que o conteúdo verificado e todos os outros conteúdos cobertos pelos documentos em vigor possam ser enviados ao Comitê para apelação.

Reduzir os danos causados pelo conteúdo que ficou ativo durante a análise detalhada. O conteúdo identificado como violador e de alta gravidade durante a primeira análise da Meta deve ser removido ou oculto até que seja feita uma análise mais detalhada. Esse conteúdo não deve permanecer na plataforma ganhando visualizações simplesmente porque a pessoa que publicou é um parceiro de negócios ou uma celebridade. Para que as decisões sejam tomadas o mais rápido possível, a Meta deve investir os recursos necessários para adequar sua capacidade de análise aos conteúdos que identifica como necessitando de análise adicional.

II. Solicitação da Meta

1. O Comitê de Supervisão tomou conhecimento da verificação cruzada, pela primeira vez, em 2021 ao decidir o caso sobre a suspensão das contas do ex-presidente dos Estados Unidos, Donald Trump. Apesar de não ter mencionado a verificação cruzada em seu encaminhamento inicial ou nos materiais enviados ao Comitê, a Meta descreveu o programa de verificação cruzada em resposta a um questionamento do Comitê sobre qualquer tratamento diferente recebido por uma conta. Como parte da decisão de maio de 2021, o Comitê fez duas recomendações relevantes sobre o programa de verificação cruzada:

- "Produzir mais informações que ajudem os usuários a entender e a avaliar o processo e os critérios de aplicação da permissão de conteúdo de valor jornalístico, inclusive como se aplica às contas de usuários influenciadores".
- "A empresa também deve explicar claramente a justificativa, as normas e os processos de análise de verificação cruzada e informar as classificações de erro das determinações tomadas pela verificação cruzada em comparação aos procedimentos comuns de aplicação de normas".

2. Em setembro de 2021, o Wall Street Journal divulgou documentação produzida pela ex-funcionária e crítica da empresa, Frances Haugen. A reportagem do Journal descreveu a verificação cruzada como uma isenção dos usuários mais influentes da Meta dos processos normais de moderação de conteúdo. O Independent informou que Frances Haugen afirmou que a empresa tinha "mentido repetidamente" ao Comitê sobre a verificação cruzada no caso Trump. A documentação interna da Meta publicada pelo Journal

revelou que alguns de seus funcionários consideravam as práticas de "whitelisting" da verificação cruzada "não defensáveis publicamente". Da mesma forma, de acordo com o Journal, os usuários que se beneficiavam do sistema de verificação cruzada na época tinham uma janela de "autocorreção" de 24 horas para editar ou remover o conteúdo violador e assim evitar quaisquer penalidades aplicadas pela Meta.

3. Em 21 de setembro de 2021, depois dos artigos do Wall Street Journal, o Comitê solicitou que a Meta se comprometesse com a transparência sobre o sistema. No dia seguinte, ela realizou um briefing com o Comitê sobre a verificação cruzada. O Comitê concluiu que "a equipe do Facebook encarregada de disponibilizar as informações não foi totalmente acessível nas respostas de verificação cruzada. Em alguns casos, o Facebook não disponibilizou informações relevantes ao Comitê, e, em outros casos, as informações fornecidas estavam incompletas.

4. Depois que o Comitê pediu maior transparência na verificação cruzada, a Meta apresentou esta solicitação de parecer consultivo sobre políticas. Após uma breve análise do sistema, a empresa descreveu a verificação cruzada como um programa que "proporciona mais níveis de análise para determinados conteúdos que nossos sistemas internos sinalizam como violadores, seja pela automação, seja pela análise humana. O objetivo é prevenir ou minimizar os erros de moderação com falsos positivos de alto risco". A Meta define falsos positivos como a remoção equivocada de conteúdo que não viola as políticas de conteúdo estabelecidas no que é permitido no Facebook e no Instagram.

5. A Meta fez as três perguntas a seguir ao Comitê:

   Devido à complexidade da moderação de conteúdo em grande escala, como o Facebook deve equilibrar seu desejo de aplicar, de forma justa e objetiva, nossos Padrões da Comunidade com a necessidade de flexibilidade, das nuances e das decisões com base no contexto de verificação cruzada?

   Quais melhorias o Facebook deve fazer na forma de regulamentar nosso sistema de verificação cruzada na Análise Secundária de Resposta Inicial para aplicar, de forma justa, nossos Padrões da Comunidade, e assim reduzir a potencial over-enforcement, mantendo a flexibilidade dos negócios e promovendo a transparência no processo de análise?

   Quais critérios devem ser usados pelo Facebook para determinar quem deve ser incluído na Análise secundária de resposta antecipada e ter prioridade como um dos diversos fatores do classificador de verificação cruzada e assim garantir a equidade no acesso a esse sistema e sua implementação?

6. O Comitê aceitou a solicitação da Meta em 21 de outubro de 2021. Após a aceitação, o Comitê enviou perguntas à Meta. O Comitê fez 74 perguntas à Meta. Delas, 58 foram respondidas totalmente, 11 respondidas parcialmente e

cinco não foram respondidas. A Meta levou meses para responder a algumas dessas perguntas.

7. O Comitê também recebeu 87 comentários públicos relacionados ao parecer consultivo: 9 da Ásia Pacífico e da Oceania, 2 da Ásia Central e do Sul, 12 da Europa, 3 da América Latina e do Caribe, 3 do Oriente Médio e da África do Norte, 3 da África Subsaariana e 55 dos Estados Unidos e do Canadá. Para ler os comentários públicos enviados sobre este parecer consultivo, clique aqui. Além disso, o Comitê realizou quatro workshops regionais focados no programa de verificação cruzada.

8. Com base na análise dessas informações, na pesquisa independente e no engajamento das partes interessadas, o Comitê responde agora às perguntas da Meta e fornece uma avaliação do sistema de verificação cruzada. A empresa também informou ao Comitê que fez mudanças significativas no programa de verificação cruzada durante o ano passado. O Comitê entende que essas mudanças são, pelo menos, em parte, um esforço para atender às críticas públicas ao programa. A explicação do programa por parte do Comitê e sua análise são baseadas em como a Meta define que o programa está funcionando atualmente. No entanto, às vezes, o Comitê faz referência ao entendimento de práticas anteriores, pois indicavam prováveis áreas de risco recorrente.

9. O Comitê analisou se o programa serve na prática para tratar e reduzir os impactos adversos de acordo com as responsabilidades de direitos humanos da Meta. Essa análise, com base nos padrões internacionais de direitos humanos e nos valores e compromissos declarados da Meta, envolve questões importantes sobre como a empresa trata seus usuários mais influentes e poderosos, permite que o conteúdo flua em suas plataformas e disponibiliza informações ao público sobre suas ações.

III. Sistema de verificação cruzada da Meta

Explicação da Meta sobre o uso da verificação cruzada

10. Os usuários do Facebook e do Instagram criam bilhões de itens de conteúdo todos os dias. A Meta está em constante moderação do conteúdo, triagem, avaliação e ação com base nas políticas de conteúdo da empresa. No Facebook, essas políticas são os Padrões da Comunidade e, no Instagram, as Diretrizes da Comunidade.

11. De acordo com a Meta, a moderação de conteúdo em grande escala apresenta desafios, e seus moderadores humanos e sistemas automatizados podem remover por engano um conteúdo que não viola as políticas da empresa. A Meta refere-se a essas decisões como falso positivo, que são uma forma de *under-enforcement* e referem-se ao conteúdo que viola as políticas da Meta, mas não determinam a violação durante a análise. A *under-enforcement* também inclui a violação de conteúdo que não é detectada por

moderadores humanos ou automatizados e opções de design do sistema que permitem que o conteúdo violado permaneça ativo após uma primeira análise.

12. O sistema de verificação cruzada trata apenas da *over-enforcement* de normas ou falsos positivos. Com esse sistema, a Meta atrasa a aplicação das normas sobre o conteúdo selecionado e identificado inicialmente como violador, a fim de permitir uma possível análise adicional para evitar falsos positivos.

13. A empresa descreveu a verificação cruzada como uma estratégia de prevenção de erros para equilibrar a proteção à voz dos usuários contra falsos positivos com a necessidade de remoção rápida do conteúdo violador. Como parte do pedido de parecer consultivo sobre políticas, a Meta destacou a inclusão de "jornalistas cobrindo zonas de conflito e líderes de comunidade a fim de aumentar a conscientização sobre casos de ódio ou violência", além de agentes cívicos onde "os usuários têm maior interesse em ver o que dizem seus líderes".

14. O sistema inclui ainda usuários que a empresa define como "parceiros de negócios", que têm pontos de contato exclusivos na Meta. Segundo a empresa, esses usuários incluem "organizações de saúde, editoras, artistas, músicos, artistas, criadores e organizações beneficentes". O Comitê entende que essa categoria inclui usuários que possam gerar lucro para a Meta, seja por meio de relações comerciais formais ou por atraírem os usuários para a plataforma e mantê-los engajados. O Comitê entende que os "parceiros de negócios" devem incluir também grandes empresas, partidos e campanhas políticas e celebridades.

15. A Meta explica ao Comitê que adiciona o termo "parceiros de negócios" para verificação cruzada a fim de evitar exclusões equivocadas que restrinjam a capacidade dos usuários e anunciantes de atingir seu público e clientes, e o impacto que tais erros podem causar na reputação ou nas finanças da empresa. Para esses usuários, a Meta quer evitar "experiências negativas tanto para os parceiros de negócios do Facebook quanto para uma quantidade significativa de seguidores".

16. A empresa afirmou que prefere a *under-enforcement* à *over-enforcement* das normas do conteúdo verificado, já que "no atual cenário comercial, considera-se que mais importante o benefício da verificação cruzada (evitar falsos positivos) do que a redução o custo dela [ou seja, visualizações de conteúdo violador], por conta da noção de censura." O Comitê interpreta essa questão como uma indicação de que, por motivos comerciais, lidar com a "percepção de censura" pode ter prioridade em relação a outras responsabilidades de direitos humanos relevantes para a moderação de conteúdo.

<u>Como funciona a verificação cruzada</u>

17. Os processos comuns de moderação de conteúdo da Meta se aplicam à maioria dos usuários. Quando o conteúdo é identificado por violar as políticas de conteúdo da Meta, a empresa aplica as normas. Assim, estão incluídas a exclusão de conteúdo e o uso de telas de aviso, dependendo do tipo de violação da política. Algumas violações também podem gerar penalidades na conta, como suspensão e término. Porém, em alguns casos, o conteúdo recebe tratamento diferenciado, como é o caso do sistema de verificação cruzada.

18. A Meta usa o termo "verificação cruzada" para se referir a um programa de prevenção de falsos positivos, que oferece camadas adicionais de análise de conteúdo antes da aplicação das normas. Durante a análise detalhada, equipes especializadas da Meta podem aplicar políticas de conteúdo somente para escalonamento. Essas políticas incluem permissões de conteúdo de valor jornalístico e espírito de política e todas as regras determinadas pela Meta exigem um contexto adicional para sua aplicação. Os processos de análise de verificação cruzada são acionados em duas circunstâncias.

19. Primeiro, a verificação cruzada oferece uma garantia adicional de **análise humana de conteúdo** feita por entidades autorizadas específicas quando houver publicação de conteúdo que exija a aplicação de normas de acordo com as políticas de conteúdo da Meta. Essa ação da empresa é denominada **Análise Secundária de Resposta Inicial** ou **ERSR**. Uma "entidade" é algo/alguém no Facebook ou no Instagram que pode publicar conteúdo, tais como páginas ou perfis do Facebook e contas do Instagram. As entidades podem representar indivíduos e grupos ou organizações. A Meta cria e mantém listas de entidades que ela decidiu ter direito a receber os benefícios fornecidos pela ERSR. Ou seja, se qualquer entidade autorizada publicar conteúdo identificado como violador dos Padrões ou Diretrizes da Comunidade, ele não será removido de acordo com os procedimentos aplicáveis a usuários comuns, mas será enviado para outros níveis de análise. Como a ERSR tem como base as listas, apenas alguns usuários pré-selecionados recebem esse benefício.

20. A segunda parte do sistema de verificação cruzada mostra uma análise adicional de alguns conteúdos identificados como violação das políticas da Meta, independente da identidade do usuário que publicou o conteúdo. A Meta denomina de **Análise Secundária Geral** ou **GSR**. Sempre que o conteúdo publicado por qualquer entidade na plataforma for considerado violador de uma política da empresa, seja por uma análise humana ou automação, a Meta usa um processo automatizado chamado "classificador de verificação cruzada". Esse processo analisa imediatamente vários fatores e determina se o conteúdo deve ser enviado para análise adicional e a prioridade em uma fila de outros conteúdos que aguardam o mesmo tipo de análise. De acordo com a Meta, o conteúdo publicado por qualquer usuário no Facebook ou Instagram pode ser selecionado para a GSR, uma vez que o sistema tem com base as características do conteúdo. A GSR foi implementada em 2021 e o Comitê entende que, até certo ponto, foi desenvolvida e implementada em toda a

plataforma em resposta às críticas à ERSR, incluindo as revelações de Haugen.

21. A detecção inicial de conteúdo nos dois tipos de verificação cruzada, que podem resultar em uma análise, pode acontecer de forma proativa, por meio de sistemas automatizados da Meta após a publicação do conteúdo, ou de forma reativa, seguindo denúncias de usuários. As ações de aplicação de normas que podem levar à análise de verificação cruzada incluem a exclusão de conteúdo e o uso de telas de aviso, dependendo do tipo de violação da política. Como a maioria das violações da política de conteúdo pode resultar em penalidades na conta, como suspensão e término, esses tipos de aplicação de normas também sofrem impactos. A verificação cruzada é aplicável ao Facebook e ao Instagram, exceto para alguns tipos de conteúdo (por exemplo, reels e podcasts) que, atualmente, não se enquadram no programa. De acordo com a Meta, "10% do conteúdo orgânico que está sujeito à aplicação da integridade não tem direito à atual análise de verificação cruzada".

22. Esse conteúdo permanece totalmente acessível na plataforma durante o período em que o conteúdo qualificado para verificação cruzada (através de GSR ou ERSR) seja identificado para aplicação e antes de estar sujeito ao processo de análise adicional, mesmo que a primeira avaliação seja de que o conteúdo viola os Padrões ou Diretrizes da Comunidade.

23. O Comitê acredita que, se a Meta tivesse mais moderadores disponíveis, os conteúdos nas filas de análise de verificação cruzada receberiam uma análise humana adicional. No entanto, a empresa optou por garantir apenas a análise humana adicional dos conteúdos que passam pela ERSR, o sistema para entidades autorizadas. A Meta não investiu os recursos necessários para que todo o conteúdo identificado por meio da GSR receba análise humana adicional. Embora os caminhos de análise para esses dois mecanismos sejam diferentes, conforme descrito abaixo, se algum moderador em qualquer etapa do processo descobrir que o conteúdo não viola as políticas da Meta, a análise termina e o conteúdo permanece ativo na plataforma.

Análise Secundária de Resposta Inicial (ERSR)

24. A Meta declara que inclui entidades nas listas ERSR, atribuindo-lhes uma "tag" que tem relação com a natureza e a sensibilidade da entidade. As tags específicas correspondem a diferentes listas de ERSR. A Meta afirma que aplica uma tag ERSR a entidades que correspondem às seguintes categorias: (1) cívicas e governamentais; (2) eventos mundiais significativos; (3) organizações de mídia, empresas, comunidades e criadores, incluindo anunciantes; (4) entidades com histórico de *over-enforcement* de normas; (5) entidades legais e regulatórias para as quais ações incorretas podem representar risco legal para a Meta, por exemplo, no contexto de litígio em andamento; (6) entidades cujo conteúdo está sob análise, ou seja, casos em que a ação de qualquer moderador prejudicaria a deliberação em andamento

ou apresentaria risco à Meta. De acordo com a empresa, além dos fatores usados para determinar se uma entidade se enquadra em qualquer uma das categorias mencionadas acima, como gastos com publicidade ou histórico de aplicação das normas, o direito ao ERSR também é determinado por uma avaliação do impacto de que um possível erro de aplicação teria sobre a Meta em termos de liderança da empresa envolvida na busca de uma solução. Em outras palavras, uma justificativa fundamental para o ERSR é evitar provocar pessoas que tenham meios de envolver executivos de nível sênior diretamente ou criar controvérsias públicas que esses executivos possam precisar remediar.

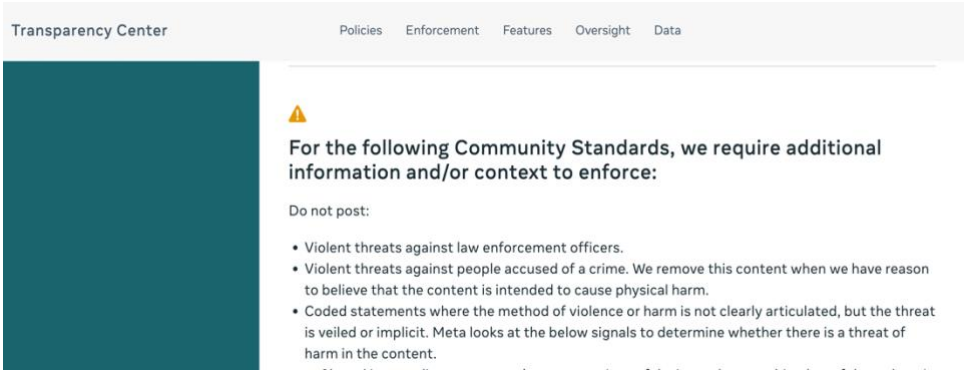
25. A Meta informou ao Comitê que está atualmente consolidando e atualizando suas listas de ERSR. Anteriormente, as listas da empresa correspondiam ao nível de escalonamento necessário para aplicar as políticas de conteúdo em relação a uma determinada entidade. Segundo a Meta, todas as entidades que atualmente têm direito à ERSR se submetem ao mesmo processo de análise. Esse processo pode incluir escalonamento opcional aos níveis mais altos da empresa.

26. A Meta informou ao Comitê que, durante o segundo trimestre de 2022, estabeleceu critérios gerais para adicionar e remover entidades das listas da ERSR e novos processos para auditorias periódicas e fiscalização interna. Ela não forneceu detalhes sobre esses processos e quais ações podem desencadear a reavaliação e a remoção de uma entidade. A Meta explicou que, em geral, as tags usadas em uma entidade de uma lista ERSR expiram após um ano e, em teoria, essas entidades precisariam ser avaliadas e marcadas novamente. Segundo a Meta, essa lógica geralmente abrange entidades nas seguintes categorias: legais e regulatórias; eventos mundiais significativos; organizações de mídia; empresas, comunidades e criadores; historicamente com *over-enforcement* das normas; e entidades escalonadas para análise de maior contexto. Meta destacou duas exceções à regra de expiração dentro de um ano. Primeiro, as tags para entidades na categoria cívica e governamental não têm expiração padrão. Segundo, as tags para entidades nas outras categorias mencionadas acima podem ter direito à ERSR mais curta a critério da Meta.

27. Quando um conteúdo de qualquer uma das entidades autorizadas é marcado para aplicação de normas, seja por análise humana ou por automação, nenhuma ação é tomada e o conteúdo é **enviado para análise detalhada por um moderador humano.** Esse primeiro nível de análise detalhada é feito pelo que a Meta chama de "**Equipe de Mercado Regional**", uma equipe interna da empresa. Essa equipe inclui funcionários da Meta e terceirizados que detêm conhecimento contextual e linguístico adicional sobre um mercado geográfico específico. Se um moderador dessa Equipe de Mercado determinar que o conteúdo não é violador, o processo será encerrado e o conteúdo permanecerá ativo na plataforma.

28. No entanto, se o moderador descobrir que o conteúdo viola as políticas da Meta, o ele permanece ativo na plataforma enquanto é escalonado para o que a Meta chama de "**Equipe de Resposta Inicial**" para outra análise. De acordo com a Meta, essa equipe tem "uma experiência da política mais detalhada e a capacidade de levar em consideração um contexto adicional".

29. A Equipe de Resposta Inicial também tem mais critérios do que outros moderadores de conteúdo da Meta e pode aplicar políticas de conteúdo que "exigem informações ou contexto adicional para serem aplicadas". A Meta geralmente marca essas políticas de conteúdo com um ponto de exclamação amarelo dentro de cada Padrão da Comunidade, conforme mostrado abaixo. Por exemplo, ao final dos Padrões da Comunidade sobre Violência e Incitação do Facebook, a Meta proíbe "ameaças violentas contra autoridades policiais". De acordo com a Meta, a determinação de manter ou remover o conteúdo que possa violar essas partes específicas do contexto da política somente pode ser feita por uma equipe com permissão para avaliar o contexto adicional, como a "**Equipe de Resposta Inicial**".



30. A **Equipe de Resposta Inicial** também pode aplicar o que a Meta denomina "permissões de conteúdo de interesse jornalístico " e "espírito da política", o que permite que o conteúdo violador permaneça ativo na plataforma, pois a empresa considera que ele é de utilidade pública ou que, mesmo que viole o texto de uma política, não viola a intenção dela. O Comitê também acredita que esse critério se estende à aplicação de penalidades na conta. No entanto, conforme divulgado pela Meta, a **Equipe de Resposta Inicial** não tem conhecimento do idioma ou região e depende das traduções e informações contextuais fornecidas pela Equipe de Mercado Regional apropriada para avaliar o conteúdo.

31. No momento dos briefings do Comitê com a Meta, aproximadamente 0,01% de todo o conteúdo identificado com necessidade de aplicação de normas, conforme uma política da Meta, foi escalonado por meio de verificação cruzada para os moderadores que podem aplicar essas políticas e permissões contextuais. Os usuários que pertencem às listas ERSR têm a garantia de que seus conteúdos publicados passarão por esses moderadores antes de qualquer ação de aplicação de normas: o conteúdo não pode ser removido ou ter uma tela de aviso aplicada por análise automatizada, por moderadores

humanos em escala ou moderadores da **Equipe de Mercado**. Durante todo o período de verificação cruzada do conteúdo até a sua determinação final, ele permanece ativo na plataforma, onde os usuários podem curtir e compartilhar livremente.

32. Depois que o conteúdo for analisado pela **Equipe de Resposta Inicial**, e se considerado violador, a Meta pode tomar as ações adequadas de aplicação de normas, como a remoção ou o uso de uma tela de aviso. No entanto, a Meta pode escalonar ainda mais a decisão. O Comitê entende que os procedimentos de escalonamento nessa fase são totalmente opcionais. Se a Equipe de Resposta Inicial decidir que o conteúdo "é uma interpretação extrema das políticas [da Meta]" ou "apresenta um risco significativo para a empresa ou a comunidade e/ou há desacordo entre as partes interessadas internas sobre a resposta", ela poderá realizar uma análise adicional em conjunto com outras equipes da Meta. De acordo com a empresa, essas análises escalonadas "incluem as opiniões e contribuições de especialistas no assunto da política de conteúdo e das equipes locais de política pública, comunicação e jurídicas" e podem incluir opiniões de outras equipes. Após essa análise, o conteúdo pode até ser escalonado para a liderança da empresa antes de receber qualquer ação de aplicação de normas.

33. Além disso, a Meta informou ao Comitê que se "a questão tiver bloqueios significativos de serviços, risco legal, regulatório ou de segurança, ou onde [houver] limites de tempo para tomar uma decisão, [a Equipe de Resposta Inicial], em raras ocasiões, escalonará a decisão diretamente à liderança sênior global." A Meta afirmou que avalia o risco de responsabilidade, urgência, impacto geopolítico, risco de bloqueio de serviço e desacordo entre equipes internas como fatores para escalonar essas decisões.

34. Em resumo, um conteúdo publicado por uma entidade de uma lista ERSR pode receber até cinco análises antes de estar sujeito à aplicação de normas, mesmo que os moderadores decidam repetidamente que o conteúdo viola as regras do Facebook ou Instagram e o escalonem para as etapas de verificação cruzada:

1. Análise inicial por automação ou moderação humana que identifica o conteúdo para aplicação das normas com base nas políticas da Meta.
2. Análise da Equipe de Mercado Regional.
3. Análise da Equipe de Resposta Inicial. Essa é a primeira análise que pode autorizar a aplicação de normas a um conteúdo. Essa equipe pode solicitar uma análise detalhada da Equipe de Resposta Inicial que incorpora outras equipes ou encaminhar diretamente para a análise da Liderança Global.
4. Análise detalhada da Equipe de Resposta Inicial com especialistas no assunto, equipes de políticas públicas, comunicações e jurídicas.
5. Análise da Liderança Global. Este é um escalonamento opcional da Equipe de Resposta Inicial com base na gravidade das consequências para a empresa.

O processo é interrompido e o conteúdo permanece na plataforma, se for considerado não violador em qualquer etapa da análise.

35. As etapas da ERSR podem levar vários dias. De acordo com a Meta, o período interno da análise de verificação cruzada pela Equipe de Mercado varia de 12 a 120 horas, dependendo da gravidade da possível violação. Na prática, a Meta afirmou que o tempo médio para a decisão final da Análise Secundária de Resposta Inicial é superior a cinco dias. Para o conteúdo publicado pelos usuários nos Estados Unidos, a Meta afirma que leva "aproximadamente 12 dias, em média, para uma decisão". Em outros países, a média de tempo de resolução é ainda mais lenta. Por exemplo, a média no Afeganistão e na Síria é de cerca de 17 dias. Nas informações fornecidas pela Meta ao Comitê, o maior tempo que um conteúdo permaneceu na fila da Análise Secundária de Resposta Inicial foi de 222 dias. A Meta forneceu vários gráficos com esses dados ao Comitê, que mostra o tempo médio de decisão entre março de 2021 e fevereiro de 2022, com a taxa de reversão e o número de trabalhos ou itens de conteúdo revisados e diferentes países

36. A empresa afirma que o conteúdo que passa mais tempo aguardando análise foi designado como de baixa gravidade em sua "estrutura de gravidade de violação". Esse esquema classifica o conteúdo com base nos Padrões da Comunidade que a primeira análise indicou violação. A estrutura da Meta classifica cada Padrão da Comunidade de acordo com o potencial dano que as violações da política podem causar, uma determinação que a empresa afirma ter feito com base em sua pesquisa. Por exemplo, ela considera o discurso de ódio mais prejudicial do que o spam, sendo o potencial discurso de ódio priorizado antes do spam na fila ERSR.

37. Assim, em setembro de 2021, o Wall Street Journal informou que o astro brasileiro do futebol Neymar publicou imagens íntimas não consensuais de outra pessoa em suas contas do Facebook e Instagram. De acordo com a denúncia do The Guardian, o vídeo permaneceu online por mais de um dia, e "uma análise interna das publicações do Neymar constatou que o vídeo foi visto 56 milhões de vezes no Facebook e no Instagram antes de ser removido", apesar de representar uma clara violação das políticas de conteúdo da Meta. De acordo com a empresa, o motivo do acesso prolongado desse conteúdo violador foi um "atraso na análise do conteúdo devido a pendências naquele momento".

38. Uma métrica principal que a Meta informou ao Comitê usada para justificar o sistema de verificação cruzada e avaliar se está funcionando bem é a "taxa de reversão". Essa é a porcentagem de conteúdo considerado não violador durante a análise de verificação cruzada, que revoga a determinação inicial e impede a aplicação de normas no conteúdo permitido pelas regras da Meta. A Meta forneceu vários números diferentes ao Comitê sobre sua taxa de reversão para conteúdo de ERSR. Segundo a Meta, em diferentes períodos no último ano, a taxa de reversão variou de 30% a 90%. Quando essa taxa é baixa, a ERSR tem mantido mais conteúdo violador na plataforma durante as

diversas camadas de análise de verificação cruzada. Quando ela é alta, a ERSR preserva mais conteúdo não-violador da aplicação equivocada de normas.

39. De acordo com a Meta, “a maioria das visualizações ocorre quando o conteúdo é novo, portanto, a celeridade na análise das decisões e na remoção do conteúdo são cruciais para evitar danos”. Portanto, o conteúdo violador, sujeito à ERSR, permanece acessível na plataforma durante todo o período em provavelmente receberá a maior parte das suas visualizações.

Análise Secundária Geral (GSR)

40. O segundo mecanismo informado pela Meta no seu sistema de verificação cruzada é a **Análise secundária geral** (GSR). Enquanto a ERSR se aplica a todos os conteúdos publicados por entidades autorizadas específicas, a GSR pode ser aplicada a qualquer conteúdo publicado na plataforma, independentemente de quem o publica, com base em uma determinação algorítmica.

41. A GSR é um sistema relativamente novo. A respeito à divulgação de informações sobre a verificação cruzada feita na denúncia de Frances Haugen durante o último trimestre de 2021, o Comitê entende que ela se referia à iteração anterior, que se baseava totalmente na entidade que publicava o conteúdo. A empresa agora também incluiu a GSR na verificação cruzada. A Meta afirmou que a GSR estava disponível para os conteúdos publicados por alguns usuários em 2021 e totalmente implementada para os conteúdos publicados por todos os usuários no início de 2022.

42. Para decidir o conteúdo enviado para GSR antes de executar uma ação de aplicação de normas, a Meta usa um algoritmo (ou seja, um conjunto de regras que precisam ser seguidas pelo computador para uma tarefa específica) chamado classificador de verificação cruzada. Esse algoritmo tem como base os seguintes fatores: “sensibilidade do tópico (qual a tendência/sensibilidade do tópico), gravidade da aplicação (a gravidade da possível medida de aplicação), probabilidade de falsos positivos, alcance previsto e sensibilidade da entidade”. A sensibilidade da entidade é, portanto, um fator nos dois sistemas, embora na ERSR seja o fator principal e na GSR seja um fator entre outros. A Meta afirmou que considera incluir fatores adicionais e espera implementá-los no futuro.

43. De acordo com a empresa, o conteúdo deve atender a duas condições para se enquadrar na GSR. Primeiro, ele precisa ter sido identificado para aplicação de normas (ou seja, violação de um Padrão ou Diretriz da Comunidade) por automação ou análise humana. Em segundo lugar, deve ser marcado pelo classificador de verificação cruzada como alta prioridade. Se ambas as condições forem atendidas, não há uma aplicação imediata sobre o conteúdo, mas ele é colocado em uma fila para análise humana adicional por uma **Equipe de Mercado Regional**. São as mesmas Equipes de Mercado

que também realizam a primeira análise detalhada de conteúdos publicados por entidades com direito à ERSR.

44. As Equipes de Mercado não conseguem analisar todo o conteúdo que tem a garantia de análise de acordo com a ERSR e todo aquele que é colocado em fila para possível análise de acordo com a GSR. Como as entidades autorizadas nas listas ERSR têm a garantia de análise, as Equipes de Mercado devem priorizar a capacidade de moderadores para esse conteúdo. Com os outros moderadores, a Equipe de Mercado analisa o conteúdo GSR identificado por algoritmos. As Equipes de Mercado também analisam alguns conteúdos fora do programa de verificação cruzada, entre outras tarefas que devem priorizar.

45. Portanto, mesmo que o conteúdo da GSR possa ter alta prioridade pelo algoritmo do classificador de verificação cruzada como merecedor de análise adicional porque pode ter sido identificado como um provável falso positivo, a Equipe de Mercado pode não conseguir revisá-lo. Em alguns casos, se não houver capacidade de análise no nível da Equipe de Mercado e a Meta tiver optado pela terceirização, alguns conteúdos da GSR podem ser enviados para essa análise adicional a moderadores humanos terceirizados. Se o conteúdo da GSR for analisado por um moderador da Equipe de Mercado, na maioria dos casos, essa decisão é final. Se for considerado violador, ele geralmente está sujeito à aplicação de normas (por exemplo, remoção ou aplicação de telas de aviso), mas, se for considerado que não é violador, o conteúdo permanece ativo na plataforma. No entanto, se a **Equipe de Resposta Inicial** tiver qualquer capacidade adicional após suas obrigações de análise de todo o conteúdo ERSR antes de sua possível remoção, essa equipe pode analisar o conteúdo GSR com alta prioridade que um moderador da Equipe de Mercado considerar violador antes que a Meta prossiga com a aplicação de normas.

46. Da mesma forma que no processo da ERSR, o conteúdo originalmente avaliado como violador dos Padrões da Comunidade e colocado na fila da GSR permanece na plataforma enquanto aguarda análise adicional. No entanto, ao contrário da ERSR, o conteúdo na fila da GSR com análise pendente não permanece ativo na plataforma de forma indefinida. O conteúdo que não é analisado periodicamente “expira” da fila da GSR. Quando isso acontece, a Meta revoga sua decisão inicial de aplicação de normas sem análise adicional. Assim, a ação que teria sido aplicada, como remoção ou tela de aviso, é feita com atraso, sem análise adicional. Se os moderadores não conseguirem analisar um conteúdo específico na fila da GSR, ele permanecerá ativo na plataforma de dois a quatro dias, antes que a Meta remova o conteúdo da fila de análise e faça a aplicação de normas. Ao mesmo tempo, o classificador de verificação cruzada identifica regularmente o conteúdo mais novo e com maior prioridade, suspende a aplicação de normas ao conteúdo e o adiciona à fila de GSR.

47. A eficácia geral da GSR fica limitada à determinação da Meta da capacidade de moderadores para esse tipo de análise em cada um de seus mercados. A

maior parte do conteúdo de GSR é analisada por um moderador terceirizado, um moderador da Equipe de Mercado ou expira no sistema. Assim, a maior parte do conteúdo da GSR nunca chega à **Equipe de Resposta Inicial** e, portanto, nunca atingirá um nível de revisão em que a análise contextual, as políticas apenas e escalonamento e as permissões de políticas possam ser aplicadas.

48. A Meta também calcula a taxa de reversão de conteúdo que recebe a análise de verificação cruzada pelas etapas da GSR. A empresa forneceu ao Comitê taxas diferentes para esse valor no ano passado. No período dos briefings do Comitê com a Meta em fevereiro de 2022, a taxa de anulação da Análise Secundária Geral era de cerca de 80%. Posteriormente, a empresa forneceu novas informações ao Comitê, afirmando que entre março e maio de 2022, a taxa de foi de cerca de 70%. Embora também houvesse diferença, esses números variaram menos ao longo do tempo. A maior parte do conteúdo de GSR, inicialmente identificado como violador, não viola nenhuma política da Meta com análise secundária. Conforme o conteúdo expira na fila de GSR, é muito provável que a Meta esteja aplicando um número significativo de falsos positivos identificados pelo classificador de verificação cruzada.

Verificação cruzada e isenções relatadas da aplicação de normas

49. A reportagem do Wall Street Journal descreveu a verificação cruzada como um sistema para isentar "usuários VIP da aplicação normal das normas da empresa". A Meta revelou ao Comitê que possui um sistema que bloqueia algumas ações de aplicação de normas fora do sistema de verificação cruzada. A Meta refere-se a essa prática como "correções técnicas" e o relatório público a descreveu como "allowlisting" e "whitelisting".

50. "Correções técnicas" são exceções automáticas à aplicação da política de conteúdo. Isso significa que substituem quase todas as tentativas de automação ou moderação humana de executar uma ação de aplicação de normas a um conjunto pré-selecionado de violações da política de conteúdo. Cada conteúdo identificado para aplicação de normas é verificado automaticamente para checar se é aplicável alguma "correção técnica".

51. Se for protegido por uma correção, o conteúdo fica isento dessa aplicação específica. Conforme explicado pela Meta, uma "correção técnica" se aplica apenas a uma entidade específica para uma violação de política específica e não serve para impedir a aplicação de normas para outras violações de política. No período dos briefings do Comitê com a Meta, a empresa afirmou que aplicava cerca de mil **correções técnicas** por dia. Ela não divulgou a quantidade e o tipo de entidades que se beneficiaram de uma "correção técnica".

52. Se o conteúdo não estiver protegido por nenhuma correção, ele será verificado quanto à elegibilidade para verificação cruzada. Nesse momento, são aplicados os processos normais de verificação cruzada da Meta para

identificar se o usuário é uma entidade que se enquadra na ERSR ou se o conteúdo é priorizado pelo classificador de verificação cruzada para a fila de GSR.

53. Primeiro, a Meta afirmou que aplica “correções técnicas” principalmente a “dois grupos de tipo de violação (spam/comportamento não autêntico e falsificação de identidade)”. Em seguida, a empresa confirmou que, em 21 de setembro de 2022, existem quatro “correções técnicas” ativas e que isso também pode mudar com o passar do tempo.

54. A Meta informou ao Comitê que “permanece um número limitado de “correções técnicas” e [Meta] reconhece uma necessidade contínua das correções”. De acordo com a empresa, essas “correções ajudam [a Meta] a evitar erros de aplicação de normas nos conteúdos ou as entidades que provavelmente não violam nossas políticas e direcionam os recursos de análise humana onde são mais necessários”.

55. A Meta reconheceu deficiências nas práticas anteriores de correções técnicas, e informou ao Comitê que a “falta de governança sobre as práticas anteriores, [...] inadvertidamente resultou em algumas entidades não receberem muitas ações de aplicação de normas”. Ela também afirmou que “diferentes equipes poderiam aplicar diferentes correções à mesma entidade de forma que, quando combinadas, a entidade e seu conteúdo não recebessem uma grande quantidade de ações de aplicação de normas”, e ainda afirmou que, como essa prática foi “o resultado inadvertido de um sistema descentralizado, [a Meta] [está] tomando medidas para garantir que haja uma estrutura de governança em torno do uso de listas de verificação cruzada”.

Verificação cruzada no contexto de solicitações do governo para remover conteúdo

56. Quando algum governo solicita que a Meta remova o conteúdo, ela pode removê-lo porque viola as políticas de conteúdo da empresa. A empresa também pode remover ou “bloquear geograficamente” o conteúdo por motivos legais, restringindo a acessibilidade em determinadas áreas. A Meta informou ao Comitê que adiciona entidades para verificação cruzada das listas ERSR para protegê-las de ações errôneas que podem apresentar risco legal para a empresa, por exemplo, no contexto de litígios em andamento.

57. De acordo com a Meta, as solicitações do governo para remover o conteúdo são tratadas por equipes especializadas que podem exigir a aplicação imediata do conteúdo, independentemente de ter sido publicado por uma entidade autorizada pela ERSR ou ter alta prioridade pelo classificador de verificação cruzada. Em outras palavras, as remoções decorrentes de solicitações do governo substituem os privilégios de verificação cruzada.

IV. Estrutura para análise do Comitê

Padrões internacionais dos direitos humanos

58. Em 16 de março de 2021, a Meta anunciou sua Política corporativa de direitos humanos, na qual destaca o compromisso de respeitar os direitos de acordo com os Princípios Orientadores sobre Empresas e Direitos Humanos da ONU (UNGPs). Os UNGPs, sancionados pelo Comitê de Direitos Humanos da ONU em 2011, estabelecem uma estrutura voluntária de responsabilidades das empresas sobre direitos humanos. Esses direitos incluem, "no mínimo, [...] aqueles expressos na Carta Internacional de Direitos Humanos" (Princípio 12).

59. Como uma empresa global comprometida com os UNGPs, a Meta deve respeitar os padrões internacionais de direitos humanos, independentemente de onde atuam, e enfrentar os impactos adversos nos direitos humanos (Princípio 11). Isso também significa que a empresa deve "buscar prevenir ou mitigar os impactos adversos nos direitos humanos que estejam diretamente relacionados às suas atividades e operações, produtos ou serviços prestados em suas relações comerciais, mesmo se elas não tiverem contribuído para esses impactos" (Princípio 13).

60. Os UNGPs também estabelecem que as empresas devem realizar uma auditoria sobre direitos humanos para avaliar os impactos reais e potenciais e agir de acordo com suas descobertas (Princípio 17). Para fazer isso de forma eficaz, as empresas devem monitorar indicadores qualitativos e quantitativos e incorporar informações das partes interessadas afetadas (Princípio 20).

61. Por meio de seus casos, o Comitê avalia os impactos sobre os direitos humanos das decisões específicas de aplicação das normas. Quando esses casos mostram que a Meta está causando um impacto negativo ou pode não estar tomando medidas para identificar, monitorar e limitar esses impactos de forma mais ampla, o Comitê faz as recomendações corretivas apropriadas. Em um parecer consultivo sobre políticas, o Comitê enfatiza diretamente as escolhas políticas da Meta, incluindo os processos de desenvolvimento e aplicação de normas para avaliar se a empresa está mantendo seu compromisso de respeitar os direitos de acordo com os UNGPs.

62. Aplicável à verificação cruzada, o Comitê analisou se o programa serve na prática para tratar e reduzir os impactos adversos nos direitos humanos de acordo com as responsabilidades da Meta. O Comitê também analisou detalhadamente as métricas usadas pela empresa para determinar a eficácia do programa e o que isso sugere em relação aos objetivos dela.

63. Em sua análise, o Comitê considera que uma grande quantidade de direitos pode ser afetada pelo programa de verificação cruzada. A liberdade de expressão, que inclui o direito de procurar e receber informações (Artigo 19, Pacto Internacional sobre Direitos Civis e Políticos Comentário Geral 34,

2011, parágrafo 11), pode ser aprimorada desde que a verificação cruzada ajude a restringir a aplicação de normas contra o conteúdo que não viola as políticas da plataforma. Assim, o resultado é um impacto positivo para o usuário da publicação e para quem deseja acessar seu conteúdo.

64. O Comitê também observa que a verificação cruzada poderia, em teoria, ajudar a garantir que aqueles que enfrentam dificuldades específicas para exercer seu direito à liberdade de expressão se beneficiem da camada de proteção adicional oferecida pelo programa. A denúncia em massa direcionada ao conteúdo não violador, por exemplo, pode ser inibida por um sistema de prevenção de erros falso-positivo.

65. No entanto, esses efeitos positivos podem ser restritos se o sistema for desenvolvido principalmente para proteger ou priorizar a expressão de pessoas que já são poderosas. O Comitê também observa que o programa de verificação cruzada levanta questões de não discriminação, pois algumas entidades recebem proteção adicional.

66. Além disso, a proteção do programa de verificação cruzada de conteúdo violador pode contribuir para um ambiente que inibe a expressão daqueles que podem ser alvo desse conteúdo violador. A variedade de conteúdo violador deixado na plataforma por mais tempo pode afetar gravemente uma variedade de direitos humanos, e as consequências podem variar dependendo da situação dos usuários afetados. Os impactos adversos nos direitos humanos provavelmente serão sentidos de forma mais aguda por indivíduos e grupos que enfrentam a marginalização e a discriminação.

67. A análise do Comitê leva em conta esses padrões. Suas recomendações de política também reconhecem as restrições da capacidade da Meta de moderar o conteúdo em grande escala. Se a moderação da empresa avaliasse com mais precisão o conteúdo de todos os usuários, não seriam necessários programas especiais com base nas entidades autorizadas para ajudar a promover o respeito aos direitos humanos.

Os valores da Meta

68. Os padrões internacionais de direitos humanos estabelecem parâmetros para as políticas e práticas da Meta. De acordo com esses padrões, no entanto, as empresas de mídia social podem adotar diferentes abordagens com relação aos direitos. Os valores da Meta devem nortear as decisões arbitrárias da empresa.

69. A Meta afirmou que possui cinco valores que influenciam o desenvolvimento da aplicação de suas políticas de conteúdo no Facebook e no Instagram. Esses valores são “Voz”, “Autenticidade”, “Privacidade”, “Segurança” e “Dignidade”. De acordo com a Meta, “Voz” é o valor “predominante” da empresa. O Comitê considera que a verificação cruzada e um sistema de

prevenção de erros falsos positivos em geral envolve principalmente “Voz”, “Privacidade”, “Segurança” e “Dignidade”.

70. Um sistema de prevenção de falsos positivos que mantém conteúdos na plataforma que não violem as políticas da Meta contribui para que o Facebook e o Instagram sejam espaços de expressão. À medida que um sistema de prevenção de erros falsos positivos mantiver um conteúdo violador e prejudicial na plataforma, facilitando o seu alcance, ele poderá impactar de modo adverso os valores “Voz”, “Segurança”, “Privacidade” e “Dignidade” de terceiros. Na medida em que o sistema privilegia a fala de alguns em detrimento de outros com o atraso e redução da probabilidade de aplicação de normas, esse tratamento desigual implica o valor de “Dignidade” da Meta, que se relaciona com a expectativa de que a empresa tratará todos os usuários de forma justa. A Meta deve garantir que seus sistemas sejam estruturados de forma a considerar todos os seus valores.

V. Avaliação do sistema de verificação cruzada

71. Nos períodos de briefings do Comitê com a Meta, ela realizava cerca de 100 milhões de tentativas de aplicação de normas nos conteúdos todos os dias. Assim, mesmo que a empresa conseguisse tomar decisões sobre a moderação de conteúdo com 99% de precisão, ainda cometeria um milhão de erros por dia. Os erros de moderação de conteúdo da Meta incluem *over-enforcement* e *under-enforcement* das normas, o que significa que ela tanto remove o conteúdo não violador como falha ao remover o conteúdo violador.

72. Nesse sentido, o uso de verificação cruzada pela Meta atende a desafios maiores na moderação de imensos volumes de conteúdo. O Comitê concorda que, nesse contexto desafiador, a Meta precisa de mecanismos para lidar com falsos positivos e falsos negativos. No entanto, a empresa tem a responsabilidade de lidar com esses grandes problemas para beneficiar todos os usuários e não apenas alguns poucos selecionados. Quaisquer decisões relacionadas ao atraso ou isenção de ações de aplicação de normas para alguns usuários ou itens de conteúdo devem estar de acordo com as responsabilidades de direitos humanos da Meta e seus valores declarados. A verificação cruzada, seja no seu formato anterior ou atual, não cumpre com essas responsabilidades e valores.

73. O Comitê observa que a Meta fez melhorias nesse sistema antes de encaminhar esta solicitação ao Comitê e durante o tempo em que ele avalia a verificação cruzada. No entanto, vários aspectos do sistema de verificação cruzada não se alinham com a responsabilidade da Meta de identificar e mitigar impactos negativos sobre os direitos humanos ou defender os valores da empresa. Isso inclui:

- Um amplo escopo para atender a objetivos múltiplos e contraditórios que permitem visibilidade e efeito viral para conteúdo violador.
- O acesso desigual à aplicação de normas e políticas arbitrárias.

- Essa inscrição no programa pode exceder a capacidade.
- A falha ao rastrear as métricas principais para avaliar o programa e fazer melhorias.
- A falta de transparência e auditabilidade sobre o seu funcionamento.

74. Apesar da grande preocupação pública com o programa, a Meta não tratou com eficácia os componentes problemáticos do seu sistema. Nesta seção, o Comitê destaca vários deles. Nas seções a seguir, faremos uma série de recomendações à Meta para destacar como um sistema de prevenção de erros poderia atender melhor os compromissos da empresa.

Amplo escopo para atender objetivos múltiplos e contraditórios que permitem a visibilidade de conteúdo violador

75. A Meta informou ao Comitê que a Revisão Secundária de Resposta Inicial existe para "proteger a voz [e] aumentar a transparência e a confiança da comunidade". A Meta ainda chamou a atenção em sua solicitação ao Comitê para a inclusão de jornalistas e líderes de comunidade na verificação cruzada. A empresa destacou que a verificação cruzada garante que a voz seja preservada em vários cenários importantes:

- "Membros de comunidades marginalizadas que recompartilham o discurso de ódio contra eles a fim de conscientizar ou condenar o conteúdo que foi removido por engano por violar as políticas de discurso de ódio".
- "Jornalistas que denunciam zonas de conflito onde as organizações designadas estão em atividade, cujo conteúdo foi removido por engano por violar nossas políticas de Organizações e Indivíduos Perigosos".
- "Nudez relacionada à saúde, como reconstrução pós-mastectomia ou fotos de amamentação removidas por engano por violar nossas políticas de nudez."

76. Em reunião com o Comitê, quando questionados sobre os impactos negativos que poderiam decorrer sem a ERSR, os funcionários da Meta afirmaram que um problema, por exemplo, seria que poderia impedir a comunicação e o fluxo de informações em caso de crise como um desastre natural ou golpe de estado. Esses pontos enfatizam a lógica declarada da Meta para o sistema que contrasta claramente com a forma como o sistema opera.

77. O Comitê compartilha a preocupação da empresa sobre a remoção indevida de conteúdo não violador publicado por pessoas dão destaque a violações de direitos humanos, trabalham para promover a saúde da mulher e outras denúncias de utilidade pública. De fato, as decisões do Comitê trataram desses erros. A Meta identifica esses casos como "erros de aplicação de normas" somente depois que o Comitê apresenta esses casos à empresa. Alguns exemplos incluem: a decisão do caso Cinto *Wampum* (2021-012-FB-UA) de remover incorretamente a expressão de combate ao ódio de um artista indígena após várias decisões errôneas de análise humana; a decisão

do Comitê sobre o caso *Menção do Talibã em uma reportagem* (2022-005-FB-UA) de remover incorretamente a publicação de uma agência de notícias sobre uma determinada organização; e a decisão do caso *Sintomas do câncer de mama e nudez* (2020-004-IG-UA) de remover incorretamente por automação uma publicação que deveria ter se beneficiado da exceção que relaciona a saúde às políticas de nudez adulta da Meta.

78. Embora, na descrição do programa, a Meta priorize as vozes em risco que publicam conteúdo não violador, ela também afirmou que o programa de verificação cruzada tem uma função comercial central, pois desempenha um "papel importante no gerenciamento de relacionamentos do Facebook com muitos [de seus] parceiros de negócios". Da mesma forma, a estrutura de sensibilidade de tags para verificação cruzada, que corrobora com o fator de "sensibilidade de entidade" para classificação GSR e tags ERSR, está diretamente ligada, entre outros fatores, ao grau de reputação e reação interna prevista se um determinado conteúdo for removido por erro. Por exemplo, a Meta caracteriza o risco de "escalonamento para níveis mais altos (CEO, COO)" como correspondente a uma tag de verificação cruzada de "gravidade extremamente alta". A correlação da prioridade mais alta na verificação cruzada com as questões sobre gerenciamento de relações comerciais sugere que as consequências que a Meta deseja evitar são principalmente relacionadas aos negócios e não aos direitos humanos.

79. A fim de avaliar como a empresa prioriza as entidades na verificação cruzada, o Comitê solicitou repetidamente que a Meta compartilhasse sua lista de Análise Secundária de Resposta Inicial para sua avaliação. Ela não disponibilizou a lista ao Comitê. O Comitê não consegue avaliar totalmente até que ponto a empresa está cumprindo suas responsabilidades de direitos humanos no programa ou o perfil das entidades que têm garantia de análise detalhada, se não entender como o programa está sendo implementado e a quem exatamente ele beneficia. A Meta afirmou que disponibilizar uma lista de usuários sujeitos à verificação cruzada violaria as obrigações legais da empresa em relação à privacidade do usuário. Com base em uma orientação jurídica, o Comitê acredita, e indicou para a Meta, que essas questões poderiam ter sido mitigadas e avisos mais detalhados fornecidos.

80. Quase cinco meses depois de ter solicitado essas informações pela primeira vez, a Meta forneceu ao Comitê uma lista com dados agregados limitados sobre cada entidade listada na atual lista de Análise Secundária de Resposta Inicial. Especificamente, a empresa divulgou apenas o tipo de entidade (por exemplo, usuário do Instagram, página do Facebook), o país e o idioma associados, conforme selecionado pela própria entidade, e se a Meta considera ou não a entidade "cívica" e um "parceiro". Nem todas as informações foram fornecidas para cada categoria da entidade. Por exemplo, um quarto das entidades listadas do Instagram não selecionou um país ou idioma específico em suas configurações de perfil e também não são consideradas um agente cívico ou parceiro da Meta.[50] Para essas entidades, a empresa apenas divulgou a sua existência, mas não as identidades ou

características de um grupo de usuários do Instagram que se beneficiam da verificação cruzada.

81. Essa divulgação limitada prejudica a capacidade do Comitê de executar suas responsabilidades de supervisão. A descrição da Meta da categoria “cívica”, por exemplo, inclui agentes estatais, representantes eleitos, “influenciadores cívicos” e candidatos a cargos públicos, entre outros. Da mesma forma, a categoria “parceiro” abrange organizações de notícias, celebridades, artistas e muito mais. O Comitê não pode avaliar, por exemplo, até que ponto jornalistas, defensores de direitos e dissidentes em determinados países recebem a mesma proteção para sua expressão que os agentes estatais inscritos na ERSR de acordo com a política do programa.

82. A Meta informou ao Comitê que não tem um sistema abrangente que avalie sistematicamente quais jornalistas, defensores de direitos humanos ou figuras da sociedade civil em um determinado local devem estar sujeitos à ERSR. A inclusão desses usuários na lista tem como base decisões descentralizadas da equipe da Meta, descritas pela empresa como “especialistas internos com grande conhecimento de mercado”. Contido, esse fato aumenta o risco de existirem lacunas e inconsistências significativas para quem recebe as camadas adicionais de proteção para expressão fornecida pela verificação cruzada da ERSR.

83. Os jornalistas que publicam conteúdo de contextos de conflito, oposição política que busca cargos públicos eletivos, celebridades que publicam uma grande variedade de conteúdo e parceiros de negócios que publicam conteúdo para vender mercadorias representam perfis de risco bastante diferentes do ponto de vista da liberdade de expressão e dos direitos humanos. De acordo com os problemas da Meta na moderação de conteúdo em grande escala, dentro das limitações atuais, o conteúdo gerado pelo usuário deve estar sujeito a diferentes prioridades focadas nos direitos. A empresa descreveu um sistema que não inclui estratégias ou táticas para garantir que os indivíduos e as expressões que mais precisam de proteção a recebam no curto prazo, com o objetivo final de oferecer uma melhor moderação de conteúdo para todos.

84. De acordo com a ERSR, caso o conteúdo de qualquer entidade autorizada seja identificado como violador e sinalizado para análise adicional, esse conteúdo, independentemente do perfil de risco, permanece na plataforma durante o pico de efeito viral após a publicação imediata. É significativo por dois motivos: primeiro, o conteúdo viral se espalha rapidamente nas plataformas. Em segundo lugar, uma vez que algo é publicado por uma entidade que tem grande alcance, o conteúdo inevitavelmente será gravado e recompartilhado individualmente pelos usuários, mesmo que a publicação original seja excluída. Isso significa que as contas que se beneficiam do sistema de verificação cruzada ERSR podem fazer upload de conteúdo violador e sabem que pode ter um grande alcance, mesmo que esteja em violação.

85. Embora o Comitê destaque que a Meta afirmou ter um sistema para priorizar o conteúdo ERSR de alta gravidade para análise, esse conteúdo ainda permanece na plataforma até que todas as análises necessárias sejam concluídas, às vezes, por períodos significativos. Por exemplo, no caso do Neymar, é difícil entender como imagens íntimas não consensuais publicadas em uma conta com mais de 100 milhões de seguidores não estariam na frente da fila para uma análise rápida e de alto nível, se havia um sistema de priorização. Por conta da gravidade da violação da política e o impacto sobre a vítima, esse caso destaca a necessidade de a Meta adotar diferentes abordagens para o conteúdo com análise pendente e reduzir os prazos de análise.

86. O atraso na aplicação das normas no conteúdo violador é uma fonte significativa de danos no programa de verificação cruzada. De acordo com a própria pesquisa da Meta, as visualizações do usuário do conteúdo violado por causa da verificação cruzada ocorrem pelas “anulações incorretas e ao atraso na aplicação de normas de não anulações para as quais a aplicação é lenta devido ao processo de análise secundária”. A empresa reconhece que a proteção adicional oferecida ao conteúdo de alguns usuários privilegiados pode gerar conflito com outros usuários pela violação de conteúdo, conforme as publicações de discurso de ódio ou de assédio.

87. O conteúdo concedido automaticamente à ERSR é diferente do conteúdo identificado e enviado para a GSR. Por um lado, conforme observado acima, a porcentagem de conteúdo ERSR considerado violador parece variar. Nos períodos em que a taxa de reversão é baixa, uma importante deficiência no sistema é a falha em garantir a remoção imediata do conteúdo violador.

88. Por outro lado, a maior parte do conteúdo GSR é geralmente considerado não violador. Para esse sistema, a taxa de reversão parece mostrar que há mais problemas na *over-enforcement* de normas em escala e que a análise secundária permite principalmente que o conteúdo não violador permaneça acessível. O Comitê observa, portanto, que, à medida em que a GSR preserva mais expressão, seu impacto é limitado pelas restrições de capacidade aplicadas pela Meta.

89. Resumindo, o Comitê considera que, apesar de a Meta caracterizar a verificação cruzada como um programa para proteger vozes vulneráveis e importantes, a verificação parece ser mais focada diretamente em atender às questões de valor comercial. Embora o Comitê entenda que a Meta é uma empresa e deve conseguir elaborar políticas que atendam às preocupações comerciais, essas mesmas políticas não devem ser caracterizadas como medidas de mitigação de riscos aos direitos humanos se não atenderem a esse objetivo. Além disso, se as escolhas de design de negócios da Meta tiverem impacto negativo nos direitos humanos, ela deve identificar e prevenir, mitigar ou terminar com esses impactos negativos por meio de melhorias no programa.

Acesso desigual à aplicação de normas e políticas arbitrárias

90. A verificação cruzada foi criada para enviar alguns conteúdos para decisões de moderação mais leves, determinando se alguma exceção ou política especializada poderia ser aplicável para recusa da aplicação de normas. De acordo com a Meta, “se o conteúdo que passou por verificação cruzada for escalonado para análise adicional, o conteúdo pode estar sujeito a uma decisão com base em [...] políticas específicas do contexto”. A verificação cruzada permite a análise humana pela “Equipe de Resposta Inicial”, na qual o Comitê acredita que pode conceder exceções na aplicação de normas, tanto relacionadas ao conteúdo específico quanto às penalidades contra a própria entidade. O conteúdo analisado pela ERSR tem a garantia de chegar a essa equipe antes de uma possível remoção, e o conteúdo analisado pela GSR tem uma chance maior de chegar a essa equipe.

91. A Meta informou repetidamente ao Comitê e ao público que o mesmo conjunto de políticas se aplica a todos os usuários. Tais declarações e as políticas de conteúdo voltadas para o público são enganosas, pois apenas um pequeno subconjunto de conteúdo chega a um moderador com poderes para aplicar o conjunto completo de políticas.

92. Portanto, o direito de análise secundária beneficia significativamente o usuário. Isso significa que a maior parte dos conteúdos que o usuário decide publicar deve permanecer ativo na plataforma. No caso de conteúdo não violador, é protegido da remoção equivocada. No caso de conteúdo violador, é permitido permanecer na plataforma durante o pico de visualizações antes de uma remoção posterior.

93. O Comitê também acredita que, além de aplicar as políticas de conteúdo com mais discrição, o conteúdo analisado no escalonamento pode se beneficiar das decisões de não aplicar restrições de conta que seriam feitas nos procedimentos normais. Em geral, as violações da política de conteúdo correspondem a “advertências” contra uma conta, que por sua vez correspondem a consequências específicas. De acordo com a Central de Transparência da Meta, as advertências resultam a períodos cada vez maiores em que as contas não podem publicar conteúdo. Em caso de advertências graves ou repetidas, a Meta desativará uma conta.

94. O Comitê questionou sobre a aplicação arbitrária das políticas e as consequências da aplicação de normas. A Meta respondeu que “não possui dados estatísticos significativos que façam distinção entre as penalidades aplicadas a entidades que passam por verificação cruzada e entidades que não passam por verificação cruzada” e “não tem conhecimento e não encontrou pesquisas ou análises” que abordem essas possíveis discrepâncias. Como a verificação cruzada pode isentar os usuários de consequências na conta, o Comitê está preocupado que a empresa tenha

optado por não rastrear e analisar essas informações ou tenha deixado de divulgá-las ao Comitê.

95. De acordo com a reportagem do The Guardian, depois que Neymar publicou o conteúdo violador, ele "não estava sujeito ao procedimento normal do Facebook para uma pessoa que publica fotos de nudez não autorizadas, que é a exclusão da conta". Esse exemplo foi apontado nas denúncias feitas pela denunciante e não ficou claro se essas práticas podem ser divulgadas. O Comitê também solicitou à Meta a confirmação das restrições no nível da conta aplicadas nesse caso. A empresa acabou revelando que a única consequência seria a remoção do conteúdo e que a penalidade normal seria a desativação da conta. O Comitê destaca que a Meta anunciou posteriormente que assinou um acordo com Neymar para que ele "transmitisse jogos exclusivamente no Facebook Gaming e compartilhasse conteúdo de vídeo com seus mais de 166 milhões de fãs no Instagram".

96. O acesso desigual a análises escalonadas, bem como exceções da política, é particularmente preocupante devido à falta de critérios objetivos ou transparentes para inclusão nas listas de Análise Secundária de Resposta Inicial. Conforme observado acima, não ficou claro como a Meta garante que aqueles usuários com maior probabilidade de *over-enforcement* ou que enfrentam desafios para exercer seus direitos à liberdade de expressão recebam essa proteção adicional. O Comitê está preocupado com o fato de que esses usuários geralmente correm maior risco, incluindo jornalistas e defensores dos direitos humanos, que podem fazer denúncias sobre organizações perigosas ou documentar abusos explícitos, são aqueles adicionados proativamente a essas listas, por conta do investimento que seria necessário para encontrar essas pessoas em todo o mundo.

97. Por outro lado, a Meta explicou ao Comitê que tem uma equipe exclusiva encarregada de garantir que todas as entidades elegíveis que representam funcionários e organizações governamentais pertencem à lista ERSR. Os critérios incluem "empresas, organizações de mídia e criadores" também parecem mais claros. De acordo com a empresa, um critério, por exemplo, é uma quantia específica de gasto ou receita gerada por uma entidade na "família de aplicativos" da Meta, embora ela possa variar ao longo do tempo.

98. O Comitê também está preocupado com o fato de que, ao operar a verificação cruzada, a Meta concentre sua atenção desproporcional em mercados mais lucrativos, em vez de se concentrar em contextos com maiores riscos aos direitos humanos, incluindo a liberdade de expressão. No momento do briefing do Comitê com a Meta, 42% do conteúdo analisado pela etapa da Análise Secundária de Resposta Inicial veio dos Estados Unidos ou Canadá. Da mesma forma, 20% de todas as entidades que constam das listas ERSR nesse momento correspondem a esses dois países. Por outro lado, de acordo com a Meta, apenas 9% das "pessoas ativas a cada mês" no Facebook eram dos Estados Unidos e Canadá. Esses dados mostram que os usuários dos Estados Unidos e do Canadá têm acesso exagerado por meio

da Análise Secundária de Resposta Inicial a formas de análise especializada que garantem acesso a todas as políticas da Meta, análise de contexto e provavelmente a possibilidade de penalidades da conta não padrão para o conteúdo violador.

99. Essa diferença está relacionada ao fato de que a “renda média por pessoa” nos EUA e no Canadá é a mais alta do mundo, cerca de três vezes maior que a Europa e cerca de 12 vezes maior que a Ásia-Pacífico. Esses fatos destacam os incentivos financeiros que determinam o funcionamento da ERSR e reforçam as questões de equidade. Por meio do programa de verificação cruzada, os usuários em mercados lucrativos com maior risco de implicações de relações públicas para Meta se beneficiam de maior direito à proteção de conteúdo e expressão do que em outros lugares.

100. Além disso, para a GSR, o classificador de verificação cruzada prioriza o conteúdo de acordo com fatores como “sensibilidade do tópico”, que potencialmente exigem avaliação automatizada do idioma do conteúdo. O Comitê está preocupado com o fato de a Meta não priorizar o treinamento de seus processos de automação em idiomas menos falados e mercados menos lucrativos. O investimento limitado na moderação desses idiomas restringe a capacidade dos algoritmos de identificar tópicos em tal conteúdo. Assim, sugere que os usuários nesses mercados, inclusive o Sul Global, podem estar em desvantagem quando avaliados em relação à elegibilidade para verificação cruzada da GSR. Da mesma forma, a Meta divulgou que “um grupo de idiomas é analisado por falantes não nativos que usam ferramentas de tradução e destaque de insultos”. Isso reforça a preocupação do Comitê de que a verificação cruzada não beneficia igualmente todos os usuários, mesmo usando a GSR.

A inclusão no programa excede a capacidade

101. A elegibilidade de ERSR e GSR excede com frequência a capacidade de análise humana que a Meta direciona para o programa de verificação cruzada. A incompatibilidade entre o volume de conteúdo designado para análise detalhada por meio desses sistemas e os recursos humanos inadequados alocados para a tarefa representa uma falha importante no sistema.

102. A Meta informou o Comitê que “não pretendia operar com um acúmulo recorrente de casos, embora as restrições de capacidade operacional e os volumes crescentes tenham levado a uma demora na Análise Secundária de Resposta Inicial. [...]Esse atraso consiste no conteúdo avaliado como provável baixa gravidade.” Não obstante a declaração da Meta de que não pretendia manter casos acumulados de forma contínua, a empresa deixou de atribuir recursos humanos suficientes para atender às necessidades de moderação de conteúdo desses programas. Além disso, conforme observado acima, nem todo conteúdo sujeito a atraso na aplicação de normas é de baixa gravidade.

103. A capacidade limitada de análise humana tem consequências diferentes, mas relacionadas, para a Análise Secundária de Resposta Inicial e a Análise Secundária Geral. Para a ERSR, a deficiência na capacidade mostra que o conteúdo permanecerá na plataforma durante o período em que deve acumular mais visualizações. Já que esse conteúdo permanece na plataforma até receber uma análise detalhada, aquele publicado por usuários ERSR com perfil de destaque e que violam as políticas da Meta permanece na plataforma durante o período de maior visualização. Embora a Meta possa primeiro tentar analisar o conteúdo que pode causar maiores danos, não ficou claro se isso é feito de forma consistente, e ainda assim reflete uma decisão de design para fornecer proteção automática a entidades selecionadas com base principalmente em critérios comerciais.

104. Para a Análise Secundária Geral, a capacidade limitada pode levar a duas consequências. Primeiro, a Equipe de Resposta Inicial pode ocupar seu tempo com o conteúdo ERSR, já que esse conteúdo deve ser analisado para executar qualquer ação de aplicação de normas. Portanto, a Equipe de Resposta Inicial geralmente não tem disponibilidade para analisar o conteúdo GSR , e o conteúdo GSR não atinge esse nível crítico de análise, em que podem ser aplicadas as políticas que exigem contexto e discrição adicionais. Em segundo lugar, a capacidade limitada no nível de análise da Equipe de Mercado mostra que muitos conteúdos GSR são removidos por padrão da fila antes mesmo da análise. Como a maior parte desse conteúdo parece sempre ser não violador, uma consequência importante da capacidade limitada para o conteúdo de Análise Secundária Geral é que a Meta remove mais conteúdo que provavelmente é não violador.

105. Essas falhas agravam as disparidades no tratamento de diferentes usuários na plataforma. Os usuários privilegiados incluídos na ERSR têm mais chances de serem analisados por um moderador, que pode aplicar o contexto para defender seu conteúdo, têm uma maior variedade de exceções de política que podem ser aplicadas para defender seu conteúdo e se beneficiam de um sistema em que mesmo o conteúdo violador tem a garantia de visualização por um período. Já os usuários comuns, cujo conteúdo pode ser acessado na análise de GSR, têm oportunidades mais limitadas de análise de seu conteúdo, costumam estar mais sujeitos às políticas de conteúdo sem análise contextual ou aplicações de exceções de política e, se o tempo limite for excedido, provavelmente terão o conteúdo não violador removido. Esse sistema tem sérias implicações nos valores de “Voz”, “Dignidade”, “Privacidade” e “Segurança” que a Meta afirma exercer.

<u>Falha no rastreamento de métricas principais para avaliar o programa e fazer melhorias</u>

106. O Comitê avaliou as métricas que a Meta usa para justificar e avaliar o programa de verificação cruzada. As métricas usadas atualmente pela empresa não englobam todas as principais preocupações e parecem não ter levado a mudanças quando deficiências foram identificadas. Além disso, a

Meta está deixando de monitorar e definir objetivos para um conjunto de métricas abrangente que mostra uma imagem completa de como operar o programa e estabelecer metas de melhoria correspondentes.

107. Conforme já mencionado, uma métrica usada pela Meta é a taxa de reversão ou a porcentagem de conteúdo que está sujeito à verificação cruzada e, por fim, não violador, apesar da identificação inicial por automação ou análise humana indicar o contrário. Segundo a Meta, “a taxa de reversão é a taxa de eficácia do sistema de verificação cruzada”. De acordo com as informações fornecidas pelo Comitê, a empresa “quer que a porcentagem [de reversão] seja alta. Se nenhuma das decisões fosse anulada por meio da verificação cruzada, significa que [a empresa] estava fazendo a verificação do conteúdo errado”.

108. Embora tenha declarado que a taxa de reversão deve ser alta, a Meta continua oferecendo as maiores proteções aos usuários da Análise Secundária de Resposta Inicial. De acordo com os números fornecidos pela empresa ao Comitê, essa taxa varia significativamente. Com essa proteção ao conteúdo sem uma alta taxa de reversão consistente, a Meta pode, de acordo com seus próprios objetivos, fazer a verificação cruzada do conteúdo errado.

109. A moderação de conteúdo em grandes volumes é marcada pela *under-enforcement*. A empresa tem como foco a prevalência de conteúdo violador como sua principal métrica pública para avaliar a eficácia de seus esforços de moderação na remoção daqueles considerados prejudiciais. Isso inclui o conteúdo que passa pelo sistema de verificação cruzada. A Meta calcula a prevalência estimando a porcentagem de todas as visualizações de conteúdo no Facebook ou Instagram, que foram visualizações de conteúdo violador. O uso da prevalência como métrica geral de sucesso pode encorajá-la a automatizar ainda mais a remoção de conteúdo e limitar a aplicação com base contextual para garantir baixa prevalência na plataforma, sem mecanismos adequados para evitar exclusões indevidas de conteúdo em escala. A alta taxa de reversão permanente na GSR, por exemplo, corrobora com essa conclusão.

110. O Comitê observa que a Meta não forneceu as informações que mostram que ela rastreia dados sobre a precisão das decisões tomadas por meio de seu sistema de verificação cruzada. Assim, embora o programa deva garantir decisões precisas de moderação de conteúdo, não parece que a empresa esteja rastreando se as decisões que passam pela verificação cruzada são mais ou menos precisas do que aquelas tomadas por meio de seus mecanismos normais de controle de qualidade em escala. Os dados de precisão seriam um indicador principal da possível influência de questões de política não relacionadas ao conteúdo nas decisões de moderação feitas na verificação cruzada. Ao medir o sucesso com base apenas na taxa de reversão, a Meta não considera se as decisões finais são corretas.

111. Além disso, a Meta afirmou que as Equipes de Mercado Regional e a Equipe de Resposta Inicial são especializadas, uma vez que têm um conjunto específico de habilidades, treinamento e acesso a ferramentas internas que permitem tomar decisões de moderação em nível de verificação cruzada. No entanto, conforme descrito acima, em alguns pontos que passam pelas análises ERSR e GSR, as decisões podem ser tomadas por moderadores terceirizados. Esses moderadores não têm o mesmo acesso ou treinamento que os funcionários da Meta. Se o objetivo do programa de verificação cruzada é elaborar decisões da política mais precisas para entidades autorizadas e conteúdo importante, deve ser um princípio básico medir a precisão das decisões de verificação cruzada em geral, mas entre determinados tipos de moderadores para entender se o design operacional está funcionando como pretendido.

112. Além disso, como o objetivo da verificação cruzada é proteger o conteúdo com maior risco de *over-enforcement* das normas, a empresa deve se concentrar em métodos adicionais para identificar esse conteúdo. A Meta divulgou que, embora esteja trabalhando ativamente para entender e mitigar *over* e *under-enforcement* de normas para usuários específicos e áreas problemáticas, ainda "precisa definir quais deles sofrem de *over-enforcement*/*under-enforcement* de normas. Enquanto aguardamos essa resolução, não temos uma forma correta de criar uma definição antes da ocorrência do fato".

Falta de transparência e auditabilidade do programa e seu funcionamento

113. Por último, o Comitê também questiona as informações limitadas que a Meta disponibilizou ao público e aos usuários sobre a verificação cruzada. Este parecer consultivo sobre políticas é resultado da falha da empresa em divulgar ao Comitê as principais informações sobre esse programa no contexto de sua deliberação sobre um caso de um usuário influente sujeito à verificação cruzada.

114. Atualmente, a Meta não informa aos usuários que estão sujeitos à ERSR, o mecanismo de verificação cruzada com base na entidade, e também não informa os usuários quando denunciam um conteúdo publicado por uma entidade verificada. A empresa também tem transparência limitada sobre como os conteúdos verificados por processos complexos de análise secundária se beneficiam.

115. Além disso, a Meta não compartilha publicamente seus procedimentos de criação de listas ERSR e sua estrutura de auditoria. O Comitê não sabe, por exemplo, se as entidades que publicam continuamente conteúdo violador são mantidas nas listas de Análise Secundária de Resposta Inicial com base no perfil. A empresa não deu nenhuma indicação de que o histórico ou frequência de violação seja um fator para a criação ou manutenção de listas de Análise Secundária. Essa falta de transparência sobre a auditoria impede que o Comitê e o público entendam todas as consequências do sistema de verificação cruzada.

Conclusões sobre a verificação cruzada

116. O Comitê reconhece que um sistema de prevenção de erros pode ser uma proteção útil contra a remoção irregular de conteúdos importantes. No entanto, se a verificação cruzada não especifica essa expressão e permite que o conteúdo violador grave permaneça na plataforma, o programa cria impactos negativos sobre os direitos humanos que a Meta não tem monitorado ou mitigado. Portanto, o Comitê conclui que a verificação cruzada atual não foi desenvolvida ou implementada a fim de atender às responsabilidades de direitos humanos da Meta e aos valores da empresa.

117. Nas decisões sobre os casos, o Comitê considera o teste em três partes no Artigo 19 do PIDCP, que avalia se as restrições à expressão atendem aos requisitos de legalidade, objetivo legítimo e necessidade e proporcionalidade.

118. A legalidade refere-se às regras que são comunicadas de forma clara e acessível. A existência, a finalidade e a natureza do sistema são indefinidas de forma que não possam ser justificadas, por conta dos efeitos significativos que a verificação cruzada tem no exercício dos direitos fundamentais. As políticas de conteúdo que são aplicadas globalmente, e que somente podem ser aplicadas com contexto adicional no escalonamento, inclusive por meio de verificação cruzada, são equivocadas.

119. O objetivo legítimo refere-se às restrições que são direcionadas aos objetivos especificados no Artigo 19, inclusive respeitar os direitos de outras pessoas e proteger a segurança nacional, a ordem e a saúde pública. As métricas pelas quais a empresa mede a eficácia de seus sistemas de aplicação de normas indicam que suas motivações estão substancialmente focadas em razões comerciais.

120. A necessidade e a proporcionalidade ponderam se as restrições à expressão são a forma menos intrusiva de atingir o objetivo legítimo. O Comitê reitera aqui suas preocupações sobre o acesso desigual aos benefícios de verificação cruzada. A Meta possui processos claros para determinar que alguns de seus usuários são entidades autorizadas, como agentes estatais e parceiros de negócios. No entanto, por não mostrar critérios claros para outros usuários que provavelmente publicarão conteúdo com valor significativo para os direitos humanos, o programa não beneficia muito outros usuários, inclusive membros de grupos marginalizados e discriminados. A empresa também não coleta e monitora informações para averiguar se esse programa gera resultados mais precisos na prática. Por fim, pela verificação cruzada, o padrão da Meta é deixar o conteúdo identificado como violador acessível em suas plataformas. Por uma questão política, a empresa ignora o que determinou ser uma resposta proporcional em escala para algum conteúdo, muitas vezes, com base apenas em questões econômicas ou de relações públicas.

121. Para cumprir as responsabilidades de direitos humanos e os valores da Meta, um sistema que impeça a *over-enforcement* de normas deve ser estruturado de forma substancialmente diferente daquela apresentada atualmente.

VI. Recomendações de aplicação de normas

122. Em resposta às questões levantadas pela Meta, o Comitê apresenta aqui as recomendações sobre os sistemas de prevenção de erros com base nas entidades e em conteúdo determinado dinamicamente. A Meta tem a responsabilidade de enfrentar os seus desafios de moderação de conteúdo para beneficiar todos os usuários, e não apenas alguns poucos selecionados. No entanto, com foco neste parecer consultivo sobre políticas, o Comitê prioriza os sistemas de prevenção de erros de escopo limitado.

Recomendações de controle do sistema de prevenção de erros com base na entidade

123. Qualquer sistema com base na elegibilidade da entidade, como a Análise Secundária de Resposta Inicial, deve ser cuidadosamente elaborado, sujeito à supervisão e monitoramento contínuo. Deve-se ter a garantia de que ele atende aos objetivos declarados e avalia os fatores externos e consequências não intencionais que possam ser causadas. Tal sistema deve proteger os usuários que provavelmente publicarão expressões que são particularmente importantes do ponto de vista dos direitos humanos.

124. É fundamental que a Meta seja clara sobre seus objetivos e adapte seus sistemas para atender estritamente a esses objetivos. Também deve evitar proteger expressões que violem as políticas de conteúdo ou compromissos de direitos humanos. Além disso, como alguns usuários podem se beneficiar de proteções adicionais e formas de expressão, a empresa deve fornecer ao público informações sobre esses processos para que possam avaliar adequadamente as informações e opiniões existentes na plataforma.

Usuários que devem ser incluídos nos sistemas de prevenção de erros com base em entidades

125. A Meta afirma que as categorias de inclusão para seu sistema de prevenção de erros com base em entidades abrangem “cívicas e governamentais”, “eventos mundiais significativos”, “mídia”, “com histórico de *over-enforcement* de normas” e “comunidades marginalizadas”, “empresas”, “ criadores de conteúdo”, “entidades escalonadas para análise” e “legais e regulamentares”.

126. Essas categorias amplas exigem classificação e especificação adicionais. Em virtude dos compromissos de direitos humanos e valores declarados da Meta, se a empresa optar por operar um sistema de prevenção de falsos positivos com base na entidade, existem certas categorias de usuários que *devem* receber essa proteção, usuários que *podem* receber essa proteção e usuários

que *não deveriam* receber tais proteções devido aos riscos que representam aos direitos humanos.

127. Em primeiro lugar, as entidades que *devem* ser incluídas são aquelas que provavelmente produzirão expressões importantes do ponto de vista dos direitos humanos, inclusive em assuntos de importância pública. Esse benefício não é apenas para esses usuários, mas também para aqueles que desejam acessar as informações compartilhadas por elas.

128. Esses usuários devem incluir, por exemplo, pessoas cujo conteúdo corre alto risco de *over-enforcement* de normas, jornalistas e agências de mídia, funcionários públicos e candidatos a cargos públicos e outros agentes cívicos, inclusive defensores dos direitos humanos e de comunidades marginalizadas. A esse respeito, o Comitê considera um sistema com base na lista como um proxy para fornecer proteções adicionais à expressão crítica, e proteção baseada simplesmente na identidade do falante. O Comitê reconhece que a Meta mantém diversas listas de entidades às quais confere maior proteção, inclusive o registro de jornalistas e sua lista de "parceiros confiáveis" da sociedade civil. Essas entidades existentes e avaliadas podem formar uma fonte a partir da qual a empresa pode formar um sistema objetivo, global e com base nos padrões de direitos humanos acessível a todos aqueles que a expressão satisfaça os critérios de inclusão.

129. Em segundo lugar, as entidades que *podem ser* incluídas podem ter como base as prioridades da empresa, usuários com valor comercial ou parceiros de negócios. Isso também pode incluir anunciantes, empresas com páginas ou grupos que correm risco de *over-enforcement* de normas, usuários que representam um risco de reputação especial para a empresa ou outros usuários que tenham relação comercial com a Meta.

130. Em terceiro lugar, existem entidades que *não devem* ser incluídas em nenhum sistema de prevenção de erros que possa atrasar toda a aplicação de normas. Estão inclusos usuários e entidades que repetidamente criam ou compartilham conteúdo que viola as políticas ou termos de serviço da empresa. O atual sistema de aplicação de normas em nível de conta da Meta, com base em advertências e penalidades, pode ser aproveitado para fins de implementação dessa regra. Caso os usuários incluídos pelo seu valor comercial publicarem conteúdos violadores com frequência, eles não devem mais continuar se beneficiando de um sistema que atrasa a aplicação das normas. A Meta tem a responsabilidade de identificar e excluí-los dos sistemas que oferecem visibilidade adicional ao conteúdo violador. Embora o número de seguidores possa ser um proxy legítimo para o grau de utilidade pública na expressão do usuário, a contagem de celebridades ou seguidores de um usuário não deve ser o único critério para um sistema de prevenção de erros com base na entidade.

131. A inclusão de todas as entidades no mesmo sistema da Meta coloca elas em competição direta por recursos limitados de análise. A empresa deve priorizar

os recursos adequados para sistemas de prevenção de erros que mitiguem os danos aos direitos humanos. Nesse contexto, a Meta deve garantir que o conteúdo com implicações de direitos humanos ou de utilidade pública seja analisado em tempo hábil por moderadores qualificados e com a capacidade de levar mais contexto em consideração, independentemente de o conteúdo ter vindo por etapas baseadas em entidade ou em conteúdo.

132. O Comitê recomenda que a empresa tome medidas para usar etapas separadas ou criar mecanismos de priorização para diferenciar os usuários que devem ser incluídos por conta das responsabilidades de direitos humanos da Meta daqueles que são incluídos devido a prioridades comerciais, oferecendo diferentes perfis de risco. As empresas, por exemplo, podem ter mais conteúdo identificado como violador das regras de spam, já que podem publicar conteúdo comercial rapidamente. Os usuários com um grande número de seguidores podem publicar sobre assuntos importantes de utilidade pública, mas também podem publicar conteúdo violador.

Os tomadores de decisão devem estar qualificados e capacitados para tomar decisões que respeitem os direitos

133. De acordo com as recomendações do Comitê, a Meta deve priorizar seus fluxos de trabalho de análise secundária de prevenção de erros de acordo com o perfil de risco e o valor dos direitos humanos.

134. O conteúdo publicado por entidades que a empresa *deve* incluir com base nas questões de direitos humanos, deve ser revisado por equipes com experiência em contexto e linguagem. Essa etapa de análise, inclusive sua trajetória de escalonamento, deve ser desprovida de valores comerciais. A Meta deve tomar medidas que assegurem que essa equipe não denuncia as equipes de políticas públicas ou de relações governamentais ou responsáveis pela gestão de relacionamento com quaisquer usuários afetados.

135. A trajetória dedicada à resolução de problemas explicitamente relacionada às prioridades comerciais da Meta poderia abordar, por exemplo, a aplicação de normas a anúncios, regras de spam, limites de recursos e problemas comportamentais. Um exemplo de problemas comportamentais é uma página de uma empresa que é injustamente penalizada por fazer upload de fotos a uma taxa muito mais rápida do que um perfil normal. Seja por ter menor prioridade ou separação em um fluxo de trabalho diferente, essas análises não devem desviar recursos direcionados à mitigação de direitos humanos.

136. O Comitê observa que a Equipe de Resposta Inicial, que tem permissão para aplicar exceções de política e interpretar o contexto, não exige que seus moderadores tenham conhecimento cultural ou linguístico. De acordo com a Meta, as decisões são tomadas com base em notas fornecidas pelas Equipes de Mercados Regionais. A própria empresa reconheceu que “não é suficiente confiar nas traduções”. Nesse contexto, o Comitê insiste que a Meta deva garantir conhecimentos culturais e linguísticos nesses níveis de análise. A

empresa deve considerar a contratação de funcionários com conhecimento cultural e linguístico de regiões de risco e o desenvolvimento de procedimentos para incluir funcionários com esse conhecimento na tomada de decisões.

Instruções para criar e controlar listas para sistemas de prevenção de erros com base em entidades

137. A Meta deve estabelecer critérios claros e públicos para elegibilidade na prevenção de erros com base na entidade. Esses critérios devem diferenciar os usuários cuja expressão merece proteção adicional do ponto de vista dos direitos humanos, inclusive informações de utilidade pública, e usuários incluídos por motivos comerciais. Por exemplo, a empresa atualmente define uma categoria de verificação cruzada como “Organizações de mídia, empresas, comunidades e criadores”. Essa categoria inclui “organizações de saúde, editoras, artistas, músicos, criadores e organizações beneficentes”. Critérios mais amplos são insuficientes. A Meta também deve desenvolver critérios com base nos padrões de violação ou comportamento indesejável na plataforma para evitar a concessão de proteções a usuários prejudiciais.

138. A Meta deve adicionar entidades aos sistemas de prevenção de erros apenas quando o processo for objetivo, bem regido e transparente. Todas as entidades propostas para serem adicionadas a uma lista devem ser informadas dessa possibilidade e devem ter a opção de recusar a inclusão, se assim o desejarem. Aquelas entidades que optarem pela inclusão devem analisar as regras de conteúdo da Meta e se comprometer novamente a segui-las. Embora o Comitê considere a verificação cruzada uma fonte de benefícios para os usuários incluídos, a Meta deve funcionar com base nos princípios de consentimento do usuário.

139. Os critérios públicos claros também devem oferecer uma base para que os usuários qualificados busquem proativamente a inclusão nessas listas. A Meta deve estabelecer um processo pelo qual os usuários possam executar proteções de prevenção de erros com *over-enforcement* de normas, caso atendam aos critérios articulados da empresa. Os agentes estatais devem ter direito a serem adicionados ou se inscreverem com base nesses critérios e termos, mas sem nenhuma outra preferência.

140. Além de atender aos critérios públicos, o processo de inclusão, independentemente de um usuário ou a Meta iniciar o processo, deve consistir em: (1) um requisito para analisar a política de conteúdo da Meta e um compromisso adicional e explícito de que seja seguido; (2) um reconhecimento das regras específicas do programa; e (3) um sistema para informar os usuários proativamente sobre mudanças nas políticas de conteúdo da empresa para facilitar a conscientização e conformidade.

141. Às vezes, a Meta trabalha com a sociedade civil por meio de seu programa de “parceiro confiável” e outras iniciativas de engajamento das partes

interessadas para coletar informações sobre entidades que devem ser consideradas para proteção. O Comitê recomenda que a Meta fortaleça seu envolvimento com a sociedade civil para criar listas. Os usuários devem poder indicar outras pessoas que atendam aos critérios públicos, desde que os indicados possam recusar a inclusão. Essa questão é particularmente urgente em países onde a presença limitada da empresa não permite identificar candidatos para inclusão de forma independente.

142. A criação de listas, e principalmente esse envolvimento, deve ser feita por equipes especializadas, independentes daquelas cujos mandatos possam gerar conflitos de interesse, como as equipes de políticas públicas da Meta. Para assegurar que os critérios serão atendidos, o pessoal especializado, com o benefício de entradas locais, deve garantir a aplicação objetiva dos critérios de inclusão. As equipes de políticas públicas, muitas vezes, interagem e fazem lobby com autoridades governamentais, criando inevitáveis incentivos conflitantes. Embora possam indicar candidatos, eles não devem ser tomadores de decisão.

143. A Meta informou o Comitê que atualmente um único funcionário da empresa pode decidir adicionar entidades a uma lista de verificação cruzada específica e não há necessidade de análise dessas decisões. No futuro, a empresa deve ter um processo estabelecido para análise objetiva e com base em critérios de todas as entidades que receberão benefícios adicionais. Pelo menos duas pessoas de equipes diferentes devem estar envolvidas para finalizar a inclusão em qualquer proteção com base na lista, e indivíduos com relações pessoais ou comerciais com entidades nomeadas não devem tomar decisões.

Instruções para manter e auditar listas para sistemas de prevenção de erros com base em entidades

144. Além de estabelecer critérios claros para a inclusão em um programa de proteção contra erros, a Meta deve estabelecer critérios e processos transparentes para auditoria e remoção. Caso as entidades não atendam mais aos critérios de elegibilidade, elas devem ser removidas.

145. A Meta informou ao Comitê que sua nova estrutura de governança proposta inclui regras para adicionar e remover entidades das listas; regras de expiração de tags; procedimentos periódicos de auditoria; e uma estrutura de supervisão. No entanto, a empresa também divulgou que havia exceções a algumas dessas regras, como entidades “cívicas e governamentais” que não têm prazos de expiração padrão. A Meta também compartilhou que está atualmente auditando um subconjunto limitado de entidades na Análise Secundária de Resposta Inicial, à medida que avança para uma estrutura de lista mais simplificada.

146. O Comitê recomenda que a Meta exija pelo menos uma análise anual de todas as entidades incluídas em qualquer sistema de prevenção de erros que forneça benefícios a tais entidades. Também deve haver protocolos

transparentes para encurtar esse período, quando necessário. Da mesma forma que as suas recomendações sobre a inclusão inicial em qualquer sistema com base na lista, o Comitê recomenda que, pelo menos, duas pessoas com estruturas de relatórios separadas participem de auditorias internas.

147. A Meta também deve garantir critérios de remoção transparentes para qualquer programa de proteção com base em lista. Um critério deve ser a quantidade de conteúdo violador publicado pela entidade. Por exemplo, pode ser baseado em uma política de "três advertências", a menos que a Meta tenha estabelecido uma penalidade mais severa para as violações em questão (por exemplo, remoção de conta de imagens íntimas não consensuais). O sistema deve fornecer avisos às entidades e, em seguida, removê-los da verificação cruzada quando acumularem sua advertência final, independentemente de a violação merecer remoção da plataforma como um todo. As entidades ter a possibilidade de recorrer da remoção e se reinscrever no futuro.

148. Por fim, o Comitê enfatiza que, embora os procedimentos de auditoria interna sejam um passo na direção certa, a auditoria interna sem supervisão externa é insuficiente. As auditorias externas, pelo Comitê ou terceiros (por exemplo, pesquisadores ou sociedade civil), são necessárias para avaliar se um sistema de prevenção de erros mitiga impactos negativos sobre os direitos humanos. Embora o Comitê reconheça sérios problemas de privacidade e segurança com a auditoria externa, ele também acredita que a Meta pode tomar medidas para anonimizar e agregar dados para lidar com essas questões.

Algumas entidades que recebem proteção adicional devem ser marcadas publicamente

149. O Comitê pediu várias vezes à Meta que informasse aos usuários e ao público sobre suas políticas e práticas. Qualquer sistema de prevenção de erros com base na entidade deve fornecer a todos os usuários da plataforma explicações claras de como a empresa aplica suas regras. Atualmente, os usuários não sabem se estão incluídos na ERSR. Além disso, os usuários que visualizam e denunciam o conteúdo publicado por usuários incluídos na ERSR não são informados de que o conteúdo pode estar sujeito a procedimentos especiais de análise.

150. O Comitê recomenda que algumas categorias de entidades protegidas pelo sistema tenham suas contas marcadas como públicas. Essas categorias incluem todos os agentes estatais e candidatos políticos, todos os parceiros de negócios, todas as entidades de mídia e todas as outras figuras públicas incluídas devido ao benefício comercial dado à empresa para evitar falsos positivos. Isso permitirá que o público responsabilize os usuários privilegiados se as entidades protegidas mantiverem o compromisso de seguir as regras e

responsabilize a Meta por aderir aos parâmetros do programa anunciados publicamente.

151. O Comitê identifica vários riscos na identificação pública de usuários inscritos em um programa de prevenção de erros falsos positivos. Primeiro, pode haver risco adverso de usuários que tentam ter o controle de contas com proteções especiais, sabendo que o conteúdo violador permanecerá na plataforma por um determinado período. Em segundo lugar, algumas categorias de usuários podem enfrentar assédio ou outros ataques se for observado que eles mantêm um relacionamento ou recebem proteção especial da empresa.

152. No entanto, o Comitê considera que esses riscos podem ser mitigados e os benefícios superam os danos potenciais. Primeiro, a Meta deve investir todos os recursos necessários para aumentar a proteção de contas para usuários sujeitos a um sistema de prevenção de erros. A empresa tem experiência em fornecer camadas extras de proteção para jornalistas e outras categorias de usuários. Tais procedimentos podem ser adotados para uso em qualquer sistema de prevenção de erros com base nas entidades. Embora o risco adversário seja real, não é intransponível nesse contexto. Apesar da lista de Revisão Secundária de Resposta Inicial não ser pública, muitos usuários já afirmam que contas com perfil de destaque estão incluídas no programa de verificação cruzada.

153. A Meta não deve identificar beneficiários que sejam defensores de direitos humanos, entidades incluídas porque estão sujeitas ao histórico de *over-enforcement* de normas e aquelas incluídas porque correm risco de danos, embora possam optar pela identificação. Os claros critérios de inclusão e separação do programa para diferentes objetivos facilitarão esse processo.

154. Por fim, quando os usuários denunciam o conteúdo publicado por uma entidade identificada publicamente como beneficiária da análise adicional, o texto da denúncia deve deixar explícito que serão aplicados procedimentos especiais, explicando as etapas e possivelmente mais tempo para resolução.

Recomendações de controle do sistema de prevenção de erros com base no conteúdo

155. Embora um sistema com base na entidade deva incluir usuários que provavelmente produzirão expressões que mereçam proteção adicional do ponto de vista dos direitos humanos e usuários que possam estar particularmente em alto risco de *over-enforcement* errônea, um sistema com base no conteúdo procura proteger esse conteúdo diretamente, sem levar em conta quem fez a publicação.

Conteúdo que deve ser selecionado e priorizado para sistemas de prevenção de erros com base em conteúdo

156. De acordo com a Meta, o sistema de resposta geral “classifica o conteúdo com base no risco de falsos positivos segundo critérios como sensibilidade ao tópico (quão sensível/popular é o tópico), a gravidade da aplicação de normas (a gravidade da possível medida de aplicação de normas), probabilidade de falsos positivos, alcance previsto e sensibilidade da entidade (com base em grande parte das listas compiladas descritas acima).

157. Os fatores de maior peso para o algoritmo de classificação são a sensibilidade do tópico e da entidade. Conforme discutido acima, a sensibilidade da entidade está, entre outros fatores, diretamente relacionada ao grau de escalonamento interno que causaria um erro. Nesse sentido, o classificador de verificação cruzada da Meta também prioriza conteúdos que possam causar danos econômicos ou de reputação, um objetivo já atendido pela ERSR. Embora a GSR possa ter sido criada para atender a algumas críticas ao antigo sistema de verificação cruzada com base exclusivamente na entidade, o que sugere que a empresa continua priorizando a expressão com base no falante e não na importância da expressão.

158. O Comitê concorda que a elegibilidade universal para um sistema de prevenção de erros falsos positivos é uma etapa importante. No entanto, tal sistema deve priorizar a identificação de conteúdo que também não é direcionado a um sistema com base na entidade. O sistema deve oferecer maior proteção com base em uma lógica de direitos humanos. Embora a Meta possa oferecer alguma proteção adicional para *over-enforcement* de normas no caso de ameaçar seus interesses comerciais, assim também aos sistemas com base em listas, não deve ser feito às custas de seus compromissos com os direitos humanos.

159. Um classificador algorítmico para um sistema de prevenção de falsos positivos poderia, por exemplo, priorizar o conteúdo com base nos tipos de decisões que são difíceis para automação e moderadores humanos em escala (por exemplo, discurso com histórico de *over-enforcement* de normas ou discurso de comunidades marginalizadas). Em conjunto, o algoritmo pode priorizar a ordem de análise desse conteúdo com base na gravidade da possível violação, na probabilidade de ser um falso positivo e na probabilidade de conteúdo viral.

160. O Comitê recomenda que, para aumentar o impacto de um sistema de prevenção de falsos positivos com base no conteúdo, a Meta considere reservar uma quantidade mínima de moderadores por equipes que possam aplicar todas as políticas de conteúdo (por exemplo, a Equipe de Resposta Inicial). Além disso, a empresa deve analisar o conteúdo que passa por análise adicional para obter informações de onde ocorrem os erros de maior impacto nos seus sistemas e dar prioridade aos recursos de análise correspondentes.

Correções técnicas

161. A Meta explicou que as “correções técnicas” proíbem totalmente qualquer aplicação de normas para uma violação de política específica em uma entidade específica. Pode haver motivos comerciais para tal proteção a um conjunto extremamente seleto de entidades, mas qualquer sistema desse tipo tem o potencial de gerar um grande risco de isentar as entidades que publicam conteúdo violador da aplicação de moderação de conteúdo. Se tal sistema for usado, ele deve estar sujeito ao mais alto nível de análise interna e externa. As “correções técnicas” isentam algumas entidades de determinadas aplicações de normas e são denominadas corretamente como uma “allowlist” ou “whitelist”, por mais limitado que seja seu escopo.

162. Todas as recomendações sobre programas de listas, tais como critérios claros e determinados, processos de análise entre equipes para conceder qualquer isenção e processos de auditoria para manter as isenções aqui aplicáveis. Além disso, deve ser proibida a isenção para o conteúdo que a Meta classifica como uma violação de alta gravidade. A Meta deve realizar auditorias periódicas para todas as ações de aplicação de normas que são bloqueadas por tais isenções. Se, conforme afirmação da própria Meta, ocorrem cerca de mil ações por dia, ela deve ter a capacidade para a sua execução. Essa auditoria, com informações sobre o escopo e a precisão do programa, deve ser incluída nos relatórios de transparência trimestrais da Meta.

163. Por fim, a empresa deve procurar de forma proativa e periódica isenções inesperadas ou não intencionais que possam persistir em iterações anteriores desse programa. Nas suas decisões, o Comitê observa a repetição de casos em que a Meta deixou de forma equivocada de atualizar ou manter os sistemas, e as consequências de tais falhas no controle de um sistema de isenção podem ser críticas.

Recomendações de controle do sistema geral de prevenção de erros

164. Além das amplas mudanças no controle de como um sistema de prevenção de erros com base em lista e conteúdo deve ser estabelecido e auditado, o Comitê também recomenda que os procedimentos desse sistema tenham como foco a mitigação de danos e estejam sujeitos ao monitoramento contínuo para aprendizado e melhoria.

Mitigação de danos após a identificação de conteúdo violador

165. Como a própria empresa reconhece, uma das principais causas de danos no sistema de prevenção de erros falsos positivos da Meta é resultado do atraso na aplicação de normas no conteúdo violador durante o período em que tem provavelmente mais visualizações. Conforme declarado acima, a própria Meta identifica que os principais fatores dos usuários visualizarem o conteúdo violador de usuários ou conteúdos verificados em suas plataformas são “revogação incorreta e atraso na aplicação de normas da não revogação para as quais a aplicação é reduzida devido ao processamento da análise

secundária”. O Comitê insiste que a Meta a tomar medidas para mitigar esses danos.

166. Primeiro, a Meta deve tomar medidas que garantam que as decisões de análise adicional serão tomadas o mais rápido possível. Devem ser feitos investimentos e mudanças estruturais com o intuito de expandir as equipes de análise. Assim, os moderadores estarão disponíveis e trabalhando nos fusos horários adequados sempre que o conteúdo for marcado para qualquer análise humana aprimorada.

167. Em segundo lugar, o Comitê recomenda que a Meta use métodos diferentes do padrão para ação de aplicação aos itens de conteúdo sujeito à análise aprimorada. Isso pode incluir o uso de meios menos intrusivos, por exemplo, rebaixamento na classificação, desaceleração do conteúdo viral, ocultação ou remoção temporária do conteúdo. Com a definição de diferentes prioridades ou etapas para o conteúdo e as entidades de natureza diferentes, deve facilitar a aplicação de diferentes consequências a diferentes tipos de conteúdo.

168. O conteúdo identificado como violador na primeira avaliação da Meta, que é de alta gravidade, por exemplo, de acordo com a estrutura da empresa, deve ser removido ou ocultado enquanto aguarda análise e não deve permanecer na plataforma acumulando visualizações simplesmente porque o usuário da publicação é um parceiro de negócios ou celebridade. A diferença entre as opções de aplicação de normas, como remoção, ocultação e rebaixamento na classificação, deve ter como base a gravidade da violação. Na teoria, a estrutura da Meta é criada para levar em consideração a probabilidade de danos a curto prazo e se o conteúdo foi identificado como provável erro de aplicação. Se o conteúdo estiver oculto por esses motivos, deve ser aplicado um aviso para os usuários indicando que a análise está pendente.

169. Em terceiro lugar, a Meta não deve operar esses programas em atraso. Ao manter o conteúdo na fila de análise que excede a capacidade, significa que o conteúdo, que pode estar em violação, permanecerá na plataforma por um longo período. O atraso de qualquer aplicação de normas enquanto leva semanas para chegar a uma decisão gera a isenção funcional das entidades verificadas em relação às regras.

170. A Meta deve investir os recursos necessários para adequar sua capacidade de análise aos conteúdos que identifica como necessitando camadas adicionais de análise. Isso não significa, no entanto, que o algoritmo deva selecionar menos conteúdo. A falha da Meta em criar uma capacidade de análise suficiente não deve atrasar a aplicação do conteúdo ou a exclusão totalmente equivocada feita por sistemas ou moderadores em escala. A empresa desenvolveu processos para priorizar a revisão e garantir que sua força de trabalho tenha um fluxo contínuo de conteúdo para análise. Já que a análise de GSR atualmente resulta em uma taxa de reversão bastante alta, o Comitê acredita que mais conteúdos se beneficiariam com essa análise.

171. Em quarto lugar, a Meta não deve priorizar automaticamente a análise secundária com base na entidade e tornar uma grande parte da análise com base em conteúdo selecionada por algoritmos dependente da capacidade extra de análise.

Garantia de disponibilidade de apelação

172. A Meta informou ao Comitê que não oferece oportunidades de apelação ou análise consistente em todos os tipos de conteúdo. As apelações de conteúdo estão sujeitas ao programa de verificação cruzada que parecem sofrer da mesma inconsistência.

173. O Comitê entende que a fazer uma apelação sobre conteúdos que já atingiram o mais alto nível de análise na empresa pode ser desnecessário, uma vez que a apelação replicaria esses caminhos. No entanto, o Comitê está preocupado porque alguns conteúdos podem não estar recebendo o direito à apelação, apesar de não atingirem os níveis mais altos. O Comitê acredita que a Meta poderia e deveria ter esclarecido melhor esse ponto quando isso foi solicitado diversas vezes por ele.

174. Particularmente, também causa preocupação a confusão no que se refere à elegibilidade de apelações para levar casos ao Comitê, tanto para os usuários restaurarem seu próprio conteúdo verificado quanto para denunciar o conteúdo de outros usuários que se beneficiam da verificação cruzada. De fato, de acordo com a Meta, “nos meses de maio e junho de 2022, uma média de 35% do conteúdo no sistema de verificação cruzada [...] não pôde ser escalonada ao Comitê de Supervisão”. Os usuários incluídos nas listas de Análise Secundária de Resposta Inicial estão entre os usuários com maior alcance na plataforma. Essa situação pode estar privando o Comitê de alguns dos casos mais críticos de moderação de conteúdo no Facebook e no Instagram.

175. Como primeiro passo, a Meta deve esclarecer sobre a elegibilidade de apelações em geral e assegurar que o conteúdo que não atingir o nível mais alto de análise possa ter apelação internamente. Em segundo lugar, a empresa deve garantir que está oferecendo uma oportunidade de apelação ao Comitê para todo o conteúdo que esteja autorizado a analisar de acordo com seus documentos administrativos, independentemente se o conteúdo atingiu os níveis mais altos de análise na Meta.

Melhoria e aprendizado

176. Para cumprir suas responsabilidades de direitos humanos, a Meta deve monitorar periodicamente as atividades que têm impacto sobre os direitos. Os resultados dessas análises devem norteá-la na realização de melhorias em

suas políticas e práticas e minimizar os danos aos direitos humanos. Nesse caso, a Meta mantém diversas métricas relacionadas ao programa de verificação cruzada que já mostram onde melhorias deveriam estar sendo feitas. O Comitê acredita que a empresa também deve disponibilizar ao público informações sobre o funcionamento desse sistema, tanto para cumprir as responsabilidades de transparência quanto para se responsabilizar por melhorias.

177. Primeiro, a Meta já mantém uma taxa de reversão para seu sistema com base nas entidades (Análise Secundária de Resposta Inicial) e para seu sistema com base nos conteúdos (Análise Secundária Geral). A Meta deve usar as tendências das taxas de reversão para informar se deve padronizar a aplicação original em um período mais curto ou se outra ação deve ser aplicada para a análise pendente. Se as taxas de reversão forem muito baixas para determinados subconjuntos de violações de políticas ou conteúdo em determinados idiomas, por exemplo, a Meta deve ajustar constantemente a celeridade e a intrusão de uma medida de aplicação de normas que deve ser feita.

178. Em segundo lugar, a empresa informou ao Comitê que realizou exercícios de análise post-mortem depois que a equipe de "avaliação de risco" da Meta identificou áreas de risco ou houve um evento que ela considerou uma falha. O Comitê recomenda que essas e outras análises sejam realizadas regularmente na verificação cruzada, com base em avaliações internas de risco que testam o sistema nos pontos principais descritos neste parecer consultivo sobre políticas.

179. Em terceiro lugar, a Meta divulgou que uma das categorias usadas para a Análise Secundária de Resposta Inicial é "entidades com histórico de *over-enforcement* das normas". Isso significa que a empresa já identificou entidades nas quais a Meta reconhece não conseguir cumprir suas políticas de forma consistente e eficaz. Além de oferecer a essas entidades acesso a programas de prevenção de erros de *over-enforcement* das normas, a empresa deve usar esses dados para informar como melhorar suas práticas de aplicação em escala. A Meta deve medir a *over-enforcement* de normas nessas entidades e deve usar esses dados para ajudar a identificar outras com o mesmo problema. A redução dessa métrica deve ser uma meta explícita e de alta prioridade para a empresa.

180. Existem métricas adicionais que a Meta deve desenvolver e monitorar para melhor alinhar as estratégias de prevenção de erros com os padrões de direitos humanos. Por exemplo, a Meta deve estabelecer novas métricas para quantificar o impacto por deixar o conteúdo violador na plataforma. Em particular, a empresa deve calcular o número de visualizações que um conteúdo que foi removido acumulou enquanto aguardava a análise por conta dos mecanismos de prevenção de erros. A Meta deve determinar uma linha base para essa métrica e relatar as metas para sua redução.

181. A empresa também divulgou que toma medidas para resolver alguns problemas relacionados à *under-enforcement* de normas. Isso inclui "classificadores para detectar conteúdo que provavelmente viola nossas políticas; relatórios de usuários que identificam o conteúdo potencialmente violador; varreduras de análise humana em que nossas equipes analisam o conteúdo potencialmente violador; Operações de Análise Inicial de Alto Risco (HERO, na sigla em inglês), um sistema em que o conteúdo que tem previsão de ser viral passa por análise humana; e denúncias de repórteres, em que os usuários que denunciam a violação de conteúdo podem apelar da decisão [da Meta]".

182. O Comitê observa que os esforços envidados fora da aplicação e apelações automatizadas em escala têm um escopo restrito. Além disso, algumas dessas iniciativas competem por recursos com a verificação cruzada. Por exemplo, a análise do HERO é feita pelas equipes de mercado, que também devem dedicar capacidade de verificação cruzada. O HERO também oferece apenas análises de conteúdo que deve se tornar viral. O Comitê concorda que o conteúdo de alto alcance pode causar mais danos, mas acredita que deve ser acompanhado por esforços para melhorar a moderação de forma abrangente. A Meta deve continuar a investir em sistemas de detecção precoce e alerta; e contratar e incorporar pessoas com conhecimento local e linguístico de sua confiança e segurança, operação de análise de conteúdo e esforços de criação de listas para sistema de prevenção de erros.

VII. Recomendações de transparência

183. O Comitê fez uma série de recomendações sobre como a Meta deve criar e administrar qualquer programa de prevenção de erros falsos positivos. As responsabilidades de direitos humanos da empresa também mostram que ela deve ser transparente para o público sobre esses programas. Os relatórios de transparência devem conter dados abrangentes para que os usuários e o público entendam como o programa funciona e quais podem ser suas consequências no discurso público. Além das métricas descritas, o Comitê recomenda que a Meta inclua:

a. Taxas de reversão para sistemas de prevenção de erros falsos positivos, desagregados de acordo com as escolhas de projeto e equipes de execução (por exemplo, Mercados, Resposta Inicial, terceirizados etc.) Por exemplo, o Comitê recomendou que a Meta crie fluxos separados para diferentes categorias de entidades ou conteúdo com base em sua expressão e perfil de risco. A taxa de reversão deve ser informada em qualquer sistema com base na entidade e no conteúdo, bem como para categorias de entidades ou de conteúdos incluídas.
b. O número total e a porcentagem de políticas de apenas escalonamento aplicadas devido a programas de prevenção de erros falsos positivos em relação ao total de decisões de aplicação.
c. O tempo médio e mediano até a decisão final para o conteúdo sujeito a programas de prevenção de erros falsos positivos, dividido por país e idioma.

d. Dados agregados sobre quaisquer listas usadas para programas de prevenção de erros, inclusive o tipo de entidade e região.
e. Taxa de remoções equivocadas (falsos positivos) em todo o conteúdo revisado, inclusive a quantidade total de danos gerados por essas remoções, medidos como as visualizações totais previstas no conteúdo (ou seja, *over-enforcement*).
f. Taxa de decisões de manutenção equivocadas (falsos negativos) sobre o conteúdo, inclusive a quantidade total de danos gerados por tais equívocos, medido como a soma das visualizações acumuladas pelo conteúdo (ou seja, *under-enforcement*).

184. O Comitê recomendou anteriormente que a Meta divulgasse as taxas de erro em geral, mas também que deveria "informar as taxas de erro relativas das determinações feitas por meio de verificação cruzada em comparação com os procedimentos comuns de aplicação de normas". O Comitê acredita que a empresa tem foco na prevalência, embora seja útil em certos contextos específicos, não oferece os incentivos certos para a empresa ou as ferramentas certas para o público entender como o ecossistema de moderação de conteúdo da Meta está funcionando.

185. A empresa informou ao Comitê que está "atualmente investindo em uma medição métrica agregada de primeira linha que ajuda a entender falsos positivos em todo o sistema e está trabalhando para formar essa métrica que esperamos compartilhar externamente em nossos relatórios de transparência. Ela seria a métrica contrária à nossa medição de falso negativo, que é atualmente informada por meio de métricas de prevalência". Isso é um passo na direção certa, e o Comitê insiste que a Meta que deve concluir esse trabalho o mais rápido possível.

186. Além das métricas enfatizadas nas seções anteriores, que servem tanto para melhorar a referência quanto para fornecer informações, a Meta deve ainda disponibilizar informações básicas em sua Central de Transparência sobre o funcionamento de qualquer sistema de prevenção de erros que identifica entidades ou usuários para proteções adicionais. O Comitê entende o potencial do adversarialismo do usuário para tentar contornar a aplicação, e a Meta pode optar por resumir alguns pontos de suas práticas de aplicação. O atual nível de transparência é inadequado e não justificado pelo medo do risco adversário.

187. De modo mais geral, o Comitê observa que fornecer mais transparência a pesquisadores externos, em particular o acesso aos dados, é um componente essencial da aplicação dos sistemas de prevenção de erros. Ao longo do engajamento das partes interessadas realizado para esta análise, o Comitê ouviu as questões sobre a Meta limitar seus atuais programas de acesso a dados para partes externas. Considerando que sistemas como a verificação cruzada exigem compensações complexas, pesquisadores independentes podem fornecer à Meta informações valiosas sobre os impactos de suas escolhas. O Comitê acredita que a Meta deve estabelecer uma trajetória para

que os pesquisadores externos obtenham acesso a dados não públicos sobre o sistema de verificação cruzada que ajude a entender o programa mais completamente por meio de investigações de utilidade pública e fornecer suas próprias recomendações para melhorias. Embora devam ser tomadas medidas de mitigação para proteger a privacidade do usuário, a Meta pode e deve permitir uma maior compreensão de como funcionam suas plataformas.

VIII. Anexo com recomendações e medidas de implementação

O Comitê fez várias recomendações à Meta em seu parecer consultivo sobre políticas. Este anexo combina essas recomendações com medidas de implementação para monitorar o progresso da empresa. A Meta deve fornecer informações sobre seu trabalho de implementação em seus relatórios trimestrais no Comitê, e deve, além disso, convocar uma reunião semestral com funcionários responsáveis de alto nível para informar o Comitê sobre seu trabalho para implementar as recomendações do parecer consultivo.

#	**Recomendação**	**Medidas de implementação**
Controle de prevenção de erros com base na entidade		
1	A Meta deve dividir, seja por caminhos distintos ou priorização, qualquer programa de prevenção de *over-enforcement* com base na lista em sistemas separados: um para proteger a expressão de acordo com as responsabilidades de direitos humanos da Meta e outro para proteger a expressão que a empresa vê como uma prioridade comercial que está fora dessa categoria.	A Meta disponibiliza ao Comitê informações detalhadas de como a inclusão e a operação são divididas para essas categorias de entidades. A Meta divulga os detalhes desses sistemas na Central de Transparência. *Monitoramento*
2	A Meta deve garantir que o percurso de análise e a estrutura de tomada de decisão para conteúdo com implicações de utilidade pública ou direitos humanos, inclusive seu percurso de escalonamento, é isento de considerações comerciais. A Meta deve tomar providências para que a equipe responsável por esse sistema não faça denúncias às equipes de políticas públicas, relações governamentais ou responsáveis pela gestão de	A Meta disponibiliza ao Comitê informações que especificam os percursos de tomada de decisão e as equipes envolvidas na moderação de conteúdo com implicações em direitos humanos ou utilidade pública. *Monitoramento*

	relacionamento com os usuários afetados.	
3	A Meta deve melhorar a forma como seu fluxo de trabalho destinado a atender às responsabilidades de direitos humanos da Meta incorpora o conhecimento contextual e linguístico na revisão detalhada, especificamente nos níveis de tomada de decisão.	A Meta disponibiliza ao Comitê informações especificando como aprimorou seu processo atual de forma a incluir conhecimento linguístico e contextual quando as decisões com base em contexto e as exceções de política estão sendo consideradas. *Monitoramento*
4	A Meta deve estabelecer critérios claros e públicos para elegibilidade na prevenção de erros com base na lista. Esses critérios devem diferenciar os usuários que merecem proteção adicional do ponto de vista dos direitos humanos e aqueles usuários incluídos por motivos comerciais.	A Meta divulga um relatório ou uma atualização da Central de Transparência que especifica os critérios para elegibilidade de análise aprimorada com base na lista para as diferentes categorias de usuários inscritos no programa. *Transparência*
5	A Meta deve estabelecer um processo para que os usuários executem proteções de prevenção de erros com *over-enforcement* de normas, caso atendam aos critérios estruturados publicamente pela empresa. Os agentes estatais devem ter direito a serem adicionados ou se inscreverem com base nesses critérios e termos, mas sem nenhuma outra preferência.	A Meta implementa um sistema de aplicação transparente e de fácil acesso para qualquer proteção contra *over-enforcement* das normas com base na lista, detalhando os objetivos do sistema e como a empresa avalia as aplicações. A Meta inclui o número de entidades que se inscreveram anualmente na prevenção de erros por meio de aplicativos, seu país e categoria na Central de Transparência. *Monitoramento*
6	A Meta deve garantir que o processo de inclusão por lista, independente de quem iniciou o processo (a própria entidade ou a Meta), abrange, pelo menos: (1) um compromisso adicional e explícito do usuário em seguir as políticas de conteúdo da Meta; (2) um reconhecimento das regras específicas do programa; e (3) um sistema pelo qual as alterações nas políticas de conteúdo da plataforma são compartilhadas proativamente com eles.	A Meta apresenta ao Comitê uma experiência completa do usuário para integração em qualquer sistema por lista, inclusive como os usuários se comprometem com a conformidade com a política de conteúdo e como são notificados sobre as mudanças na política. *Monitoramento*
7	A Meta deve fortalecer seu envolvimento com a sociedade	A Meta disponibiliza informações ao Comitê sobre como a empresa se

	civil para criar listas e nomeações. Os usuários e as organizações da sociedade civil de confiança devem poder indicar outras pessoas que atendam aos critérios. Essa questão é particularmente urgente em países onde a presença limitada da empresa não permite identificar candidatos para inclusão de forma independente.	envolve com a sociedade civil para determinar a elegibilidade por lista. A Meta fornece dados na Central de Transparência, divididos por país, sobre as entidades que são adicionadas como resultado do envolvimento da sociedade civil em oposição à seleção proativa da Meta. *Monitoramento*
8	A Meta deve utilizar equipes especializadas, independentes de influência política ou econômica, inclusive equipes de políticas públicas da Meta, para avaliar as entidades que serão incluídas na lista. Para assegurar que os critérios serão atendidos, a equipe especializada, com o benefício de entrada local, deve garantir a aplicação objetiva dos critérios de inclusão.	A Meta disponibiliza ao Comitê documentos internos que especificam as equipes que lidam com a criação de listas e onde estão posicionadas na organização. *Monitoramento*
9	A Meta deve exigir que mais de um funcionário esteja envolvido no processo final de adição de novas entidades a qualquer lista para os sistemas de prevenção de erros falsos positivos. Essas pessoas devem trabalhar em equipes diferentes, mas relacionadas.	A Meta disponibiliza ao Comitê informações especificando o processo pelo qual novas entidades são adicionadas às listas, inclusive quantos funcionários devem aprovar a inclusão e a quais equipes pertencem. *Monitoramento*
10	A Meta deve estabelecer critérios claros para remoção. Um critério deve ser a quantidade de conteúdo violador publicado pela entidade. As desqualificações devem ser baseadas em um sistema de advertências transparente, no qual os usuários são avisados de que a violação recorrente pode levar à remoção do sistema e/ou das plataformas da Meta. Os usuários devem poder recorrer a tais advertências por meio de um processo justo e de fácil acesso.	A Meta disponibiliza ao Comitê informações especificando o limite de ações de aplicação de normas contra entidades nas quais a proteção de acordo com um programa com base na lista é revogado, inclusive notificações enviadas aos usuários quando recebem advertências contra sua elegibilidade, quando são desqualificados e suas opções de apelação. A Mera também deve fornecer ao Comitê dados sobre a quantidade de entidades removidas a cada ano por publicar conteúdo violador. *Monitoramento*

11	A Meta deve estabelecer critérios e processos claros para auditoria. Caso as entidades não atendam mais aos critérios de elegibilidade, elas devem ser removidas imediatamente do sistema. A Meta deve analisar todas as entidades incluídas em qualquer sistema de prevenção de erros pelo menos uma vez por ano. Também deve haver protocolos transparentes para encurtar esse período, quando necessário.	A Meta fornece ao Comitê dados sobre o valor, o tipo de entidade e o motivo da remoção das listas de entidades como resultado de auditorias, junto com um cronograma para a realização de auditorias periódicas. *Monitoramento*
Transparência nas listas		
12	A Meta deve marcar publicamente as páginas e contas de entidades que recebem proteção com base na lista nas seguintes categorias: todos os agentes estatais e candidatos políticos, todos os parceiros de negócios, todos os agentes de mídia e outras figuras públicas incluídas por conta do benefício comercial dado à empresa para evitar falsos positivos. Outras categorias de usuários podem optar por serem identificadas.	A Meta marca todas as entidades nessas categorias como beneficiárias de um programa de prevenção de erros com base na entidade e anuncia a mudança na Central de Transparência. *Transparência*
13	A Meta deve notificar os usuários que denunciam o conteúdo publicado por uma entidade identificada publicamente como beneficiária da análise adicional que serão aplicados procedimentos especiais, explicando as etapas e possivelmente uma resolução mais demorada.	A Meta fornece ao Comitê notificações sobre os usuários que denunciam conteúdo de usuários identificados como beneficiados por revisão adicional e confirmam a implementação global e os dados que mostram que essas notificações são mostradas com frequência aos usuários. *Monitoramento*
14	A Meta deve notificar todas as entidades incluídas nas listas para receber uma análise detalhada e permitir que recusem a inclusão.	A Meta fornece ao Comitê (1) as notificações enviadas aos usuários, informando sobre sua inclusão em um programa de análise aprimorada com base na lista e oferecendo a opção de recusa; e a Meta (2) divulga publicamente os números anuais na Central de Transparência sobre a quantidade de entidades, por país, que recusaram a inclusão.

		Monitoramento
Análise aprimorada e priorização		
15	A Meta deve reservar um número mínimo de moderadores por equipes que possam aplicar todas as políticas de conteúdo (por exemplo, a Equipe de Resposta Inicial) para analisar o conteúdo marcado pelos sistemas de prevenção de erros com base no conteúdo.	A Meta fornece ao Comitê a documentação que mostra o processo de avaliação dessa recomendação e a justificativa para sua decisão de implementá-la e a publicação da justificativa na Central de Transparência. *Monitoramento*
16	A Meta deve tomar medidas que garantam que as decisões de análise adicional para sistemas de prevenção de erro, que geram atraso na aplicação de normas, serão feitas o mais rápido possível. Devem ser feitos investimentos e mudanças estruturais com o intuito de expandir as equipes de análise. Assim, os moderadores estarão disponíveis e trabalhando nos fusos horários adequados sempre que o conteúdo for marcado para qualquer análise humana aprimorada.	A Meta fornece ao Comitê dados que demonstram uma redução trimestral no tempo de decisão para todo o conteúdo que recebe análise aprimorada, dividido por categoria para inclusão e por país. *Monitoramento*
17	A Meta não deve atrasar todas as ações no conteúdo identificado como potencialmente violador e deve explorar a aplicação de intersticiais ou remoções pendentes de qualquer análise aprimorada. A diferença entre remoção ou ocultação e rebaixamento deve-se ter como base uma avaliação de danos e pode ser baseada, por exemplo, na política de conteúdo que possivelmente foi violada. Se o conteúdo estiver oculto por esses motivos, deve ser aplicado um aviso para os usuários indicando que a análise está pendente.	A Meta atualiza sua Central de Transparência com a nova abordagem para ação de aplicação de normas durante o período em que o conteúdo recebe análise aprimorada e fornece ao Comitê informações detalhando a consequência da aplicação de normas aplicadas com base em critérios específicos de conteúdo. A Meta compartilha com o Comitê os dados sobre a aplicação destas medidas e o seu impacto *Monitoramento*
Recursos		

18	Em terceiro lugar, a Meta não deve operar esses programas com atraso. A Meta não deve, no entanto, obter ganhos com a capacidade relativa de análise, aumentando artificialmente o limite do classificador ou selecionando seu algoritmo sem o conteúdo.	A Meta fornece ao Comitê dados que demonstram uma redução trimestral no conteúdo total acumulado e na quantidade de dias com atraso para filas de verificação cruzada. *Monitoramento*
19	A Meta não deve priorizar automaticamente a análise secundária com base na entidade e selecionar uma grande parte da análise com base no conteúdo por algoritmos dependendo da capacidade extra de análise.	A Meta fornece ao Comitê documentos internos detalhando a distribuição de tempo e volume de análise entre sistemas com base em entidade e com base em conteúdo. *Monitoramento*
20	A Meta deve garantir que o conteúdo recebido de qualquer tipo de análise aprimorada por ser importante do ponto de vista dos direitos humanos, inclusive conteúdo de importância pública, seja analisado pelas equipes que podem aplicar exceções e contexto.	A Meta disponibiliza ao Comitê informações que mostram a porcentagem de conteúdo que recebe análise por equipes que podem aplicar exceções e contexto, porque foi publicado por uma entidade autorizada ou porque foi identificado por algoritmo como merecedor de análise aprimorada dividida por sistema de prevenção de erros (por exemplo, GSR x ERSR). *Monitoramento*
Restrições automáticas para aplicação de normas (“correções técnicas”)		
21	A Meta deve estabelecer critérios claros para a aplicação de quaisquer restrições automáticas à aplicação de normas (“correções técnicas”) e não permitir que essas restrições para violações de política de conteúdo de alta gravidade. Pelo menos duas equipes com estruturas de relatórios separadas devem participar da concessão de correções técnicas para permitir a verificação entre equipes.	A Meta publica anualmente o número de entidades que atualmente se beneficiam de uma “correção técnica”, com indicação de quais políticas de conteúdo não devem ser aplicadas. *Monitoramento*
22	A Meta deve realizar auditorias periódicas para garantir que as entidades se beneficiam de restrições automáticas à execução (“correções	A Meta presta informações ao Comitê sobre seus processos periódicos de auditoria de listas. *Monitoramento*

	técnicas") atendam a todos os critérios de inclusão. Pelo menos duas equipes com estruturas de relatórios separadas devem participar dessas auditorias para a verificação entre equipes.	
23	A Meta deve conduzir auditorias periódicas de várias equipes para procurar proativamente e periodicamente por restrições inesperadas ou não intencionais à aplicação que possam resultar de erros do sistema.	A Meta publica informações anualmente sobre quaisquer restrições inesperadas encontradas, seu impacto e as medidas tomadas para solucionar a causa-raiz. *Monitoramento*
Justeza procedimental		
24	A Meta deve garantir que todo o conteúdo que não atinja o nível mais alto de análise interna possa ser apelado à Meta.	A Meta publica informações sobre a quantidade de decisões de conteúdo feitas por meio de formas de análise aprimorada que não podem fazer uma apelação. Esses dados anuais, divididos por país, devem ser separados de forma que explique, se houver, a porcentagem do conteúdo que não recebeu apelação porque atingiu a análise de liderança global. *Monitoramento*
25	A Meta deve garantir que está oferecendo uma oportunidade de apelação ao Comitê para todo o conteúdo que esteja autorizado a analisar de acordo com seus documentos administrativos, independente se o conteúdo atingiu os níveis mais altos de análise na empresa.	A Meta confirma publicamente que todo o conteúdo coberto pelos documentos administrativos do Comitê recebe IDs de apelação do Comitê de Supervisão para enviar uma reclamação ao Comitê, fornecendo documentação para demonstrar onde foram tomadas medidas para fechar as lacunas da disponibilidade de apelação. A Meta cria um canal acessível para os usuários obterem reparação imediata quando não recebem um ID de apelação do Comitê de Supervisão. *Monitoramento*
Melhoria e aprendizado		
26	A Meta deve usar os dados que compila para identificar "entidades com histórico de *over-enforcement* das normas" e informar como melhorar suas práticas de aplicação em escala. A Meta deve medir a	A Meta fornece dados ao público que mostram quedas trimestrais na *over-enforcement* e a documentação que mostra que a análise de conteúdo de entidades "com histórico de *over-enforcement*" está sendo usada para

	over-enforcement de normas nessas entidades e deve usar esses dados para ajudar a identificar outras com o mesmo problema. A redução dessa *over-enforcement* deve ser uma meta explícita e de alta prioridade para a empresa.	reduzir as taxas *over-enforcement* de forma mais geral. *Monitoramento*
27	A Meta deve usar as tendências das taxas de reversão para informar se deve padronizar a aplicação original em um período mais curto ou se outra ação deve ser aplicada para a análise pendente. Se as taxas de reversão forem muito baixas para determinados subconjuntos de violações de políticas ou conteúdo em determinados idiomas, por exemplo, a Meta deve ajustar constantemente a celeridade e a intrusão de uma medida de aplicação de normas que deve ser feita.	A Meta fornece ao Comitê dados que detalham as taxas nas quais o conteúdo em fila permanece ativo ou é retirado, dividido por país, área de política e outras métricas relevantes, e descreve as alterações feitas anualmente. *Monitoramento*
Melhoria na responsabilidade do programa		
28	A Meta deve realizar análises periódicas de diferentes aspectos de seu sistema de análise aprimorado, inclusive conteúdo com maior tempo de resolução e conteúdo violador de alto perfil que permanece na plataforma.	A Meta publica anualmente os resultados das análises do sistema de verificação cruzada, inclusive resumos das alterações feitas como resultado dessas análises. *Transparência*
29	A Meta deve informar publicamente as métricas que quantificam os efeitos adversos no atraso da aplicação como resultado de sistemas de análise aprimorada, como visualizações acumuladas no conteúdo preservado na plataforma como resultado de sistemas de prevenção de erros, mas posteriormente considerado violador. Como parte de seus relatórios públicos, a Meta deve determinar uma linha base	A Meta inclui uma ou mais métricas importantes que demonstram as consequências negativas do atraso na aplicação, pendente de mecanismos de análise aprimorados no Relatório de Aplicação de Padrões da Comunidade, juntamente com as metas para reduzir essas métricas e o progresso no cumprimento dessas metas. *Transparência*

	para essas métricas e informar as metas para reduzi-las.	
30	A Meta deve publicar relatórios regulares de transparência focados especificamente em sistemas de prevenção de falsos positivos com atraso na aplicação. Os relatórios devem conter dados que permitam aos usuários e ao público entender como funcionam esses programas e quais podem ser suas consequências no discurso público. O Comitê recomenda que a Meta inclua, pelo menos: a. Taxas de reversão para sistemas de prevenção de erros falso positivos, divididos de acordo com diferentes fatores. Por exemplo, o Comitê recomendou que a Meta crie fluxos separados para diferentes categorias de entidades ou conteúdo com base em sua expressão e perfil de risco. A taxa de reversão deve ser informada em qualquer sistema com base na entidade e no conteúdo, bem como para categorias de entidades ou de conteúdos incluídas. b. O número total e a porcentagem de políticas de apenas escalonamento aplicadas devido a programas de prevenção de erros de falsos positivos em relação ao total de decisões de aplicação de normas. c. O tempo médio e mediano até a decisão final para o conteúdo sujeito a programas de prevenção de erros falsos positivos, dividido por país e idioma. d. Dados agregados sobre quaisquer listas usadas para	A Meta lança relatórios anuais de transparência, inclusive essas métricas. *Transparência*

	programas de prevenção de erros, inclusive o tipo de entidade e região. e. Taxa de remoções equivocadas (falsos positivos) em todo o conteúdo analisado, inclusive a quantidade total de danos gerados por essas remoções, medidos como visualizações totais previstas no conteúdo (ou seja, *over-enforcement*). f. Taxa de decisões de manutenção equivocadas (falsos negativos) sobre o conteúdo, inclusive a quantidade total de danos gerados por esses equívocos, medido como a soma das visualizações acumuladas pelo conteúdo (ou seja, *under-enforcement*).	
31	A Meta deve fornecer informações básicas na Central de Transparência sobre o funcionamento de qualquer sistema de prevenção de erros que usa para identificar entidades ou usuários para proteções adicionais.	Foi adicionada uma seção à Central de Transparência explicando sua variedade de sistemas de prevenção de erros (o Comitê entende o potencial do adversarialismo do usuário para tentar contornar a aplicação, e a Meta pode optar por resumir alguns pontos de suas práticas de aplicação). *Transparência*
32	A Meta deve estabelecer uma trajetória para que os pesquisadores externos obtenham acesso a dados não públicos sobre os programas de prevenção de erros de falsos positivos que ajude a entender melhor o programa por meio de investigações de utilidade pública e fornecer suas próprias recomendações para melhorias. O Comitê entende que as preocupações com a privacidade de dados devem exigir verificação rigorosa e agregação de dados.	A Meta divulga uma forma para que os pesquisadores externos obtenham dados não públicos sobre programas de prevenção de erros falsos positivos. *Transparência*