IGUAIS PERANTE AS PLATAFORMAS?

EQUIDADE E TRANSPARÊNCIA NA MODERAÇÃO DE CONTEÚDO EM PLATAFORMAS DIGITAIS

Diagnósticos & Recomendações #9
Julho 2023

Francisco Brito Cruz (coord.),
Alice Lana e Iná Jost

INTERNETLAB

IGUAIS PERANTE AS PLATAFORMAS?
EQUIDADE E TRANSPARÊNCIA NA MODERAÇÃO DE CONTEÚDO EM PLATAFORMAS DIGITAIS

JULHO 2023

COMO CITAR ESTE DOCUMENTO
BRITO CRUZ, Francisco (coord.), LANA, Alice de Perdigão; JOST, Iná. "Iguais perante as plataformas? Equidade e Transparência na Moderação de Conteúdo em Plataformas Digitais". São Paulo: InternetLab, 2023.

EQUIPE DO PROJETO
AUTORES Francisco Brito Cruz (coord.), Alice Lana e Iná Jost
COLABORAÇÃO Heloisa Massaro e Laura Matta
COMUNICAÇÃO João Vitor Araújo
PROJETO GRÁFICO Gabriela Rocha
APOIO Global Network Initiative

INTERNETLAB

GLOBAL NETWORK INITIATIVE

ÍNDICE

SOBRE O INTERNETLAB

O Internetlab é um centro independente de pesquisa interdisciplinar que promove o debate acadêmico e a produção de conhecimento nas áreas de direito e tecnologia, sobretudo no campo da Internet. Constituído como uma entidade sem fins lucrativos, o InternetLab atua como ponto de articulação entre acadêmicas/os e representantes dos setores público, privado e da sociedade civil, incentivando o desenvolvimento de projetos que abordem os desafios de elaboração e implementação de políticas públicas em novas tecnologias, como privacidade, liberdade de expressão e questões ligadas a gênero e identidade.

OBJETIVOS

Este projeto de pesquisa do InternetLab tem como objetivo contribuir para o debate público sobre a moderação de conteúdo em plataformas digitais. Buscamos desvendar o funcionamento de sistemas de moderação em camadas, que trazem etapas adicionais de análise qualificada para determinados tipos de perfis ou conteúdos ao decidir quais postagens devem permanecer ou ser removidas das plataformas.

Baseamos nosso estudo em algumas questões, como:

• As políticas de moderação de conteúdo das plataformas devem contemplar camadas adicionais de análise para diferentes tipos de perfis ou conteúdo?

• Se certas pessoas ou conteúdos forem tratados de maneira diferente pelas plataformas e por seus processos de moderação de conteúdo, qual estrutura deve ser usada para garantir a eficiência e a legitimidade desses sistemas?

• Como esses sistemas devem ser projetados para proteger os direitos dos usuários, especialmente no que diz respeito à equidade e transparência?

O objetivo deste documento é apresentar o conceito e as nuances dos sistemas de moderação em camadas e emitir recomendações de sistemas alinhados com os direitos humanos, a equidade e a transparência. Isso em oposição àqueles que criem bolhas privilegiadas ou listas VIP para que determinados usuários desfrutem de diferentes padrões ao publicar conteúdo online, por exemplo, devido a incentivos exclusivamente corporativos.

INTRODUÇÃO

Em 2019, o jogador de futebol Neymar publicou em suas contas do Facebook e Instagram imagens íntimas não consensuais de uma mulher, compartilhadas em uma conversa privada. As postagens faziam parte da estratégia que o atleta planejou para responder publicamente a uma acusação de estupro. Embora as políticas da Meta proíbam a publicação de imagens íntimas não consensuais, o conteúdo permaneceu na plataforma por mais de 24 horas, sendo visto por cerca de 56 milhões de pessoas.

O episódio envolvendo Neymar exemplifica um *modus operandi* da plataforma que seria confirmado dois anos depois. Em setembro de 2021, o Wall Street Journal publicou uma matéria revelando a existência de um sistema desenvolvido pela Meta que adicionava uma camada adicional ao processo de moderação de conteúdo em suas plataformas. O mecanismo, chamado de *programa Cross-check* pela empresa, fornece um exame diferente para usuários específicos, como políticas/os eleitas/os, parceiros comerciais importantes da Meta, usuários com alto número de seguidores, entre outros. Na prática, quando perfis que pertencem à lista postam conteúdo marcado como potencialmente infrator, suas postagens são direcionadas para uma fila diferente, supervisionada por uma equipe especializada, em vez da equipe de moderação regular.

Uma analogia útil é a fila de embarque no aeroporto. Todos concordam que pessoas idosas ou com bebês devem embarcar primeiro. Mas e se a fila, na prática, se aplicasse principalmente a "clientes *premium*"?

A divulgação do *Cross-check* da Meta levantou várias questões em relação à justificação e legitimidade de tais sistemas. A implementação desses mecanismos suscita preocupações sobre transparência, tratamento igualitário e riscos aos direitos fundamentais. Deveria existir uma moderação em camadas baseada em listas de usuários? Elas distorceriam ou promoveriam a equidade e a transparência na operação das plataformas? Se elas produzem algum efeito positivo, quais seriam os melhores parâmetros para que possam ser implementadas?

MODERAÇÃO DE CONTEÚDO “EM CAMADAS”: CONCEITO E CASOS

LIBERDADE DE EXPRESSÃO E SISTEMAS DE MODERAÇÃO DE CONTEÚDO

Antes de aprofundar o entendimento sobre as características de um sistema de moderação em camadas como o *Cross-check* da Meta, é importante dar um passo para trás e estabelecer um consenso sobre conceitos e definições.

Como uma definição operacional utilizada pelo InternetLab em nossa abordagem sobre o assunto, moderação de conteúdo refere-se a uma atividade-chave para uma plataforma digital: elaborar e aplicar regras, procedimentos e sistemas para remover, limitar o alcance, rotular conteúdo, suspender ou remover contas[1], bem como *“sistemas e regras das plataformas que determinam como elas tratam o conteúdo gerado pelo usuário em seus serviços”*[2]. Esse exercício corresponde à gestão das expressões individuais de um usuário, assim como à parte do produto e valor que as plataformas oferecem.

A atividade de moderação de conteúdo representa um desafio logístico para as plataformas, uma vez que lidam com uma imensa quantidade de conteúdo e contextos diversos e complexos. Isso está bem estabelecido na literatura que aborda seus principais desafios e é um ponto defendido por estudiosas/os de diferentes perspectivas. Há pesquisadoras/es que consideram que a inteligência artificial poderia apresentar uma

1 Thiago Dias Oliva, Victor Pavarin Tavares e Mariana G. Valente, “Uma solução única para toda a internet? Riscos do debate regulatório brasileiro para a operação de plataformas de conhecimento”, Diagnósticos & Recomendações (São Paulo: InternetLab, 2020). Pg. 11 Disponível em: https://internetlab.org.br/wp-content/uploads/2020/09/policy_plataformas-conhecimento_20200910.pdf

2 Doeuk, Evelyn. Content Moderation as Systems Thinking. (Harvard Law Review, 2022). Pg. 528. Disponível em: https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/

resposta eficaz para a escala massiva de dados e o constante estado de violações. Outros defendem a existência de uma estrutura de tomada de decisão sistemática, que vai além da lógica de avaliações individuais, buscando evitar a incapacidade da operação dos serviços de moderação[3].

Ainda em busca de soluções alternativas para a gestão do discurso em massa[4], sistemas de moderação em camadas poderiam ser uma estratégia empregada pelas empresas para mitigar riscos aos direitos humanos, uma vez que dão prioridade de análise a alguns tipos de usuários ou conteúdos que devem ser cuidadosamente revisados visando proteger certos tipos de discurso. Faz sentido, por exemplo, que ativistas ou jornalistas tenham sua expressão avaliada com mais cuidado do que usuários comuns, pois suas palavras têm um alcance e impacto diferentes perante o público, e suas contas e discursos podem ser constantemente alvo de ataques estratégicos por oponentes ou antagonistas.

Por exemplo, as contas e discursos publicados por defensores dos direitos humanos e jornalistas tendem a ser - potencialmente mais do que para outros atores da sociedade civil - alvo de ataques e assédio, que podem se traduzir efetivamente em intimidação e tentativas de silenciar suas vozes. Às vezes, esses tipos de ameaças podem até representar riscos significativos para sua segurança e bem-estar. Portanto, proteger seu discurso e fornecer contas com priorização de análise pode ser uma abordagem interessante para promover mais segurança.

Em outras palavras, idealmente, a moderação em camadas pode ser uma ferramenta que cria equidade dentro de um processo de moderação de conteúdo em larga escala, funcionando como uma tentativa de mitigar distorções criadas pelo processo regular e industrial de moderação pelas plataformas.

Mas e se a moderação em camadas servir apenas para preservar parceiros de negócios e interesses comerciais? E se as regras do sistema não forem claras e a sua engrenagem acabar por promover mais desigualdades, ao contrário da proteção dos direitos humanos?

3 Ibid, pg. 551.

Gillespie, Tarleton. Content moderation, AI, and the question of scale. (Big Data & Society, 2020). Pg. 2-4. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/2053951720943234

Gillespie, Tarleton. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. (Yale University Press, 2018). Disponível em: https://yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/

Klonick, Kate. The new governors: the people, rules, and processes governing online speech. (Harvard Law Review, 2017). Disponível em: https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf

Suzor, Nicolas. Lawless. The secret rules that govern our digital lives. (Cambridge, University Press, 2019). Disponível em: https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE

4 O termo "moderação em camadas" é empregado para tratar de um tipo de moderação de conteúdo que prevê uma diferença no tratamento pela plataforma a depender do usuário ou do conteúdo. Essa diferença contempla outras camadas de verificação de conteúdo que podem acrescentar, por exemplo, um estágio de análise humana para determinados casos. O que discutimos neste documento é se o sistema deve existir e como ele deve ser projetado para proteger o discurso em vez de proteger interesses que não estão comprometidos com a liberdade de expressão.

MODERAÇÃO EM CAMADAS NA PRÁTICA: O EXEMPLO DO *CROSS-CHECK*

A existência de sistemas que oferecem tratamento diferenciado a alguns usuários certamente não é exclusiva da Meta, mas, como mencionado anteriormente, a reportagem publicada pelo *Wall Street Journal* em 2021 revelou detalhes importantes desse programa, assim como a lacuna na indústria em relação à transparência desses sistemas[5].

O sistema implementa níveis privilegiados de análise para contas específicas - que a Meta determina como *"especialmente suscetíveis ao risco de sofrer ações resultantes de falsos positivos"*[6] - com base em critérios como o tipo de usuário ou entidade (político, jornalista, parceiro comercial importante, organização de direitos humanos), número de seguidores ou tópicos abordados pela entidade. Para reduzir a discricionariedade, apenas um grupo seleto de funcionários da Meta pode adicionar novas entidades à lista, que é auditada regularmente.

Quando usuários que pertencem à lista especial têm um conteúdo marcado como potencialmente infrator, eles são direcionados para a fila do *Cross-check* em vez da moderação regular[7]. Os critérios de priorização para análise dessas peças de conteúdo são *"sensibilidade do tópico (quão popular/sensível é o tópico); severidade da aplicação (a gravidade da ação de aplicação potencial); probabilidade de falsos positivos, alcance previsto e sensibilidade da entidade"*[8].

Após a divulgação, em outubro de 2021, o Comitê de Supervisão da Meta (OSB, sigla em inglês para OverSight Board) aceitou um pedido da empresa para revisar o *Cross-check* e fazer recomendações para sua melhoria. Um ano depois, o órgão divulgou um parecer consultivo trazendo conclusões importantes e orientações para aprimorar o sistema[9]. Em termos gerais, o OSB concluiu que, ao fornecer tratamento desigual para alguns usuários, o *Cross-check*: (i) causou um atraso na remoção de conteúdo violador postado pelos usuários da lista; (ii) deixou de rastrear e divulgar as métricas empregadas pelo sistema; (iii) carecia de transparência em relação ao seu funcionamento. De acordo com o Comitê, *"enquanto existem critérios claros para inclusão de parceiros de negócios e líderes governamentais, usuários cujo conteúdo provavelmente é importante do ponto de vista dos direitos humanos, como jornalistas e organizações da sociedade civil, têm caminhos menos claros para acessar o programa"*.

5 Horwitz, Jeff. Wall Street Journal. "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt". Disponível em: https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

6 A Meta define os falsos positivos como a remoção equivocada de conteúdo que não viola as políticas de conteúdo que estabelecem o que é permitido no Facebook e no Instagram. Pg. 6. Disponível em: https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

7 Toda a operação do *Cross-check* está detalhada no parecer consultivo (Policy Advisory Opinion) emitido pelo Comitê de Supervisão. Pg. 9-21. Disponível em: https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

8 Ibid. Pg. 19.

9 O OSB recebeu 87 comentários públicos relacionados a esse parecer consultivo de política: nove da Ásia-Pacífico e Oceania, dois da Ásia Central e do Sul, 12 da Europa, três da América Latina e Caribe, três do Oriente Médio e Norte da África, três da África Subsaariana e 55 dos Estados Unidos e Canadá. Disponível em: https://internetlab.org.br/wp-content/uploads/2023/05/Public-comments-appendix.pdf

Entre outras recomendações, o Comitê sugeriu que a empresa priorizasse as formas de expressão que são fundamentais para os direitos humanos, além de aumentar a transparência em torno do funcionamento do *Cross-check* e das medidas de redução de danos causados pelo conteúdo mantido durante o processo de moderação em camadas - que tende a gerar atrasos de remoção. Um resumo das 32 recomendações que o Comitê de Supervisão da Meta publicou sobre o programa em seu parecer de orientação política pode ser encontrado abaixo.

Que questões foram apresentadas pela Meta ao Comitê de Supervisão?	**1.** "Devido à complexidade da moderação de conteúdo em grande escala, como o Facebook deve equilibrar o desejo de aplicar de forma justa e objetiva nossos Padrões da Comunidade com a necessidade de flexibilidade, sutileza e decisões dependentes de contexto dentro do cross-check?" **2.** "Que melhorias o Facebook deve fazer na forma como governamos nosso sistema de cross-check de "análise secundária de resposta inicial" (ERSR), para aplicar de forma justa nossos Padrões da Comunidade, minimizando o potencial de aplicação excessiva, mantendo flexibilidade comercial e promovendo transparência no processo de revisão?" **3.** "Que critérios o Facebook deve usar para determinar quem é incluído na revisão secundária de ERSR e priorizado como um de muitos fatores pelo nosso classificador cross-check, a fim de garantir equidade no acesso a este sistema e sua implementação?"

PRIMEIRO EIXO: CONSIDERAÇÕES SOBRE DIREITOS HUMANOS E INTERESSE PÚBLICO (APLICAÇÃO)		
Tipo	**Recomendação**	**Justificativa**
Dar prioridade à expressão dos direitos humanos/ interesse público	Inclusão de usuários suscetíveis de produzir expressões importantes para os direitos humanos ou de interesse público especial na lista de prioridades do *X-Check*. Separação destes usuários dos parceiros de negócios da Meta (ou prioridades comerciais) incluídos na lista. Garantia de que o percurso e a estrutura de tomada de decisões para este conteúdo não leve em conta considerações comerciais.	Evitar a concorrência direta por recursos limitados de revisão da Meta.
Processo de inclusão	Informar os membros de que foram incluídos na lista e fornecer-lhes opções de exclusão (*opt-out*), se desejarem. Exigir que os convidados revisem as regras de conteúdo da Meta e se comprometam a segui-las antes de serem adicionados ao *X-Check*. Exigir o conhecimento específico das regras do programa. Desenvolver um sistema para informar proativamente os usuários sobre as alterações às políticas de conteúdo da Meta, de modo a facilitar o conhecimento e a conformidade.	O *X-Check* é visto como um benefício, ou tratamento desigual, para os usuários incluídos. A Meta deve operar com base nos princípios de consentimento do usuário, transparência e equidade.

Processo de inclusão	Envolver-se com a sociedade civil para fins de criação de listas.	Ter uma perspectiva multissetorial (multi-*stakeholder*) em sistemas de moderação privilegiados.
Critérios baseados em conteúdo	Desenvolver critérios baseados em conteúdo para proteger diretamente o conteúdo com alto risco de aplicação excessiva errônea (*erroneous over-enforcement*), independentemente de quem o publicou.	A abordagem atual baseada em entidades é insuficiente para garantir que conteúdos importantes de interesse público e de direitos humanos (que podem vir de qualquer usuário) não sejam removidos.
Sistema baseado em direitos humanos	Desenvolver um segundo sistema de proteção, com foco na detecção de falsos positivos (conteúdo removido erroneamente) causados pelo *X-Check* e com base em uma perspectiva de direitos humanos. Priorizar a ordem de revisão desse conteúdo com base na gravidade da possível violação, na probabilidade de ser um falso positivo e na probabilidade de viralização.	Um classificador algorítmico para um sistema de prevenção de falsos positivos poderia priorizar o conteúdo com base nos tipos de decisões mais difíceis para a moderação automática e para moderadores humanos em escala (por exemplo, discurso historicamente sujeito a aplicação excessiva ou discurso de populações historicamente marginalizadas).
Especia-lização da equipe	Criar equipes especializadas para a criação de listas para garantir que os critérios sejam atendidos, contando com contribuições locais. As equipes de políticas públicas podem indicar candidatas/os, mas não podem ser os responsáveis pela decisão final. Indivíduos com relações pessoais ou comerciais com entidades indicadas não devem ser tomadores de decisão.	Reduzir o conflito de interesses com outras equipes, como as equipes de políticas públicas da Meta, que frequentemente interagem com atores governamentais e de *lobby*. Garantir a aplicação objetiva dos critérios de inclusão.
Auditoria do *X-Check* e remoção	Promover a revisão anual de todas as entidades incluídas em qualquer sistema de prevenção de erros que ofereça benefícios a essas entidades.	Manter um padrão de elegibilidade para o sistema *X-Check*.
SEGUNDO EIXO: CONSIDERAÇÕES SOBRE TRANSPARÊNCIA		
Tipo	**Recomendação**	**Justificativa**
Transpa-rência e aplicação	Estabelecer critérios claros e públicos para inclusão no *X-Check*. Permitir que usuários que atendam a esses critérios solicitem a própria inclusão no *X-Check*.	Permitir que usuários solicitem proteções contra aplicação excessiva do *X-Check* caso atendam aos critérios estabelecidos pela empresa.

Transparência radical	Incluir nos relatórios de transparência: a. Taxas de reversão de sistemas de prevenção de erros de falsos positivos, discriminadas de acordo com diferentes fatores. Publicar taxas de reversão para sistemas baseados em entidades e em conteúdo, e categorias de entidades ou conteúdo incluídas. b. O número total e a porcentagem de políticas de escalonamento aplicadas devido a falsos positivos no *X-Check* em relação ao total de decisões de aplicação. c. Tempo médio e mediano para a decisão final do *X-Check*, discriminados por país e idioma. d. Dados agregados sobre quaisquer listas usadas para o *X-Check*, incluindo o tipo de entidade e região. e. Taxa de remoções errôneas (falsos positivos) em relação a todo o conteúdo revisado, incluindo o total de danos gerados por esses falsos positivos medidos como perda do total previsto de visualizações no conteúdo (ou seja, aplicação excessiva). f. Taxa de decisões incorretas de manutenção (falsos negativos) sobre o conteúdo, incluindo o total de danos gerados por esses falsos positivos, medido como a soma das visualizações acumuladas no conteúdo (ou seja, subaplicação).	Terceiros poderão avaliar se o programa está funcionando de forma eficaz.
Divulgação dos usuários	Marcar publicamente contas de algumas categorias de entidades protegidas pelo *X-Check* (por exemplo, atores estatais, candidatas/os, políticas/os e parceiros de negócios).	Permitir que terceiros responsabilizem usuários privilegiados por manterem o compromisso com as regras.
Priorizar a expressão dos direitos humanos/ interesse público	Nunca divulgar beneficiários que são defensores dos direitos humanos. Fornecer a eles a opção de identificação pública. Utilizar os dados compilados pela Meta para identificar "entidades historicamente sujeitas a aplicação excessiva".	Evitar danos decorrentes de aplicação excessiva histórica.
Direitos de apelação	Garantir que o conteúdo analisado pelo *X-Check* possa ser apelado para o Comitê de Supervisão, quando aplicável, independentemente de o conteúdo ter chegado aos mais altos níveis de revisão dentro da Meta.	Oferecer uma rota alternativa para apelações de aplicação indevida do *X-Check*.

Aprimoramento do *X-Check*	Publicar relatórios sobre métricas dos efeitos adversos da aplicação tardia (ou seja, divulgar visualizações acumuladas em conteúdo violador que foi mantido devido ao *X-Check*). Estabelecer uma linha de base para essas métricas e relatar metas para reduzi-las.	Indicadores de erro devem ajudar a Meta e terceiros a encontrar soluções para aumentar as remoções corretas de conteúdo no futuro ou questionar a expansão do sistema.
Informações para pesquisadores	Criar um canal no qual pesquisadoras/es obtenham dados não-públicos e anonimizados sobre o *X-Check* para investigações de interesse público e forneçam recomendações para melhorias.	Pesquisadoras/es especializadas/os podem avaliar se o programa está funcionando de forma eficaz e contribuir para sua melhoria.
Auditorias de terceiros	Promover auditorias externas, pelo Comitê de Supervisão ou terceiros (por exemplo, pesquisadoras/es ou sociedade civil) com dados anonimizados e agregados.	Avaliar se um sistema de prevenção de erros mitiga impactos negativos em direitos humanos.
TERCEIRO EIXO: REDUÇÃO DE DANOS CAUSADOS PELO CONTEÚDO REMANESCENTE DA ANÁLISE APRIMORADA		
Tipo	**Recomendação**	**Justificativa**
Penalidades alternativas	Considerar alternativas à remoção, como diminuir a visibilidade, reduzir a viralidade, ocultar ou remover temporariamente as postagens.	Reduzir os danos causados pela remoção imediata de conteúdo potencialmente violador.
Priorizar a expressão de direitos humanos/ interesse público	Permitir que revisores realizem uma análise cultural e linguística dos textos, levando em consideração contextos nacionais, regionais ou locais. Dar a revisores treinados a capacidade de levar em consideração informações adicionais sobre o contexto, independentemente de o processo de revisão ser baseado em entidades ou em conteúdo.	A Equipe de Resposta Antecipada não exige que seus revisores possuam expertise cultural ou linguística (mesmo em regiões de alto risco).
Taxas de reversão	Utilizar a taxa de reversão de decisões para determinar se deve ser mantida a aplicação original em um prazo mais curto ou qual outra medida de aplicação deve ser tomada durante a revisão.	Tomar decisões de revisão com base na taxa de erro (taxas de reversão). Se os erros forem consistentemente baixos para certas violações de política ou certos idiomas, a Meta precisa ajustar a rapidez e a intrusividade das medidas de moderação de conteúdo.

PESQUISA PARA ENQUADRAR E AVALIAR A MODERAÇÃO DE CONTEÚDO EM CAMADAS

O exercício da moderação de conteúdo é fundamental para o funcionamento das plataformas e apresenta muitos aspectos que abrem caminhos para a pesquisa, especialmente por causa de seu impacto na circulação do discurso. No início de 2022, o InternetLab iniciou uma pesquisa sobre sistemas em camadas na moderação de conteúdo, buscando criar estruturas para ajudar a avaliar se tal sistema é necessário, bem como seus limites, mecanismos, garantias e salvaguardas para os direitos humanos. Se a ferramenta é importante para lidar com os desafios logísticos da moderação e até mesmo com outras questões politicamente sensíveis, como ela deve ser projetada para não representar riscos significativos aos direitos fundamentais e, na verdade, promover os direitos humanos?

Além disso, nossa pesquisa tinha um interesse específico. Ademais de entender a necessidade dos sistemas e discutir parâmetros de transparência, queríamos utilizar uma perspectiva regional para aprofundar as vantagens e desvantagens de sua aplicação em contextos sociais, políticos, econômicos e culturais específicos, como por exemplo, em países da América Latina.

Realizamos uma série de grupos focais com partes interessadas da América Latina cujas opiniões sobre moderação de conteúdo seriam úteis. Nosso objetivo principal era identificar as questões centrais levantadas pelos sistemas de moderação em camadas a partir de perspectivas diversas e discutir alternativas de políticas para construir diretrizes sustentáveis. O material foi compilado e as principais conclusões estão expostas abaixo. Após aprofundar essas descobertas, dividimos nossas conclusões em duas perspectivas: a visão otimista e a visão pessimista, ou seja, a abordagem do copo meio cheio e a do copo meio vazio.

PESQUISA EM GRUPOS FOCAIS: MÉTODO E CONCLUSÕES

Inicialmente, foram realizados dois grupos focais com diferentes tipos de partes interessadas (stakeholders). Os participantes foram selecionados em diversos setores com presença no ambiente online, levando em consideração aspectos de classe, gênero,

raça e LGBTQIA+, buscando a equidade. Ambas as reuniões foram realizadas sob a regra de Chatham House, para garantir que todas as pessoas se sentissem à vontade para falar livremente. A regra garante que o conteúdo do que foi dito no encontro poderá ser usado pelas/os participantes apenas mediante a exclusão de informações de identificação pessoal ou relacionadas às entidades representadas.

No primeiro grupo, convidamos sete pessoas da América Latina que estudam ou atuam nos campos de integridade eleitoral, desinformação e jornalismo. Também convidamos pessoas que identificamos como influenciadores no ambiente online. O segundo grupo também foi composto por sete pessoas, também da região, que estudam ou atuam nos campos de direitos digitais, tanto da academia quanto da sociedade civil. As duas sessões foram divididas em três etapas: (i) compartilhamento de experiências individuais; (ii) questionamentos sobre sistemas de moderação em camadas; (iii) propostas para o futuro - guiadas pelas seguintes perguntas:

Compartilhamento de experiências individuais	Experiências (vividas ou observadas) sobre moderação de conteúdo, especialmente falsos negativos e falsos positivos.
	Como foi a resposta da plataforma? Ela atrapalhou ou ajudou? Como poderia ter sido melhor, considerando a quantidade de moderação que deve ser feita diariamente?
Questionamentos sobre o sistema	É necessário um sistema semelhante ao *Cross-check*? Para o quê/quem?
	Isso aumenta ou prejudica a proteção da liberdade de expressão e de outros direitos humanos?
Propostas para o futuro	Quais critérios devem definir o tipo de conteúdo a ser verificado de maneira cruzada - ou *Cross-check* - (ex.: alcance, assunto, qualquer outro)?
	Quais critérios devem definir quais contas devem ser alvo de verificação cruzada - ou *Cross-checked* - (ex.: número de seguidores, assunto, qualquer outro)?
	Como e por quem esses critérios devem ser definidos e atualizados?

Após as sessões, todas as pessoas participantes foram convidadas a apresentar contribuições por escrito sobre suas percepções em relação aos riscos e legitimidade desses sistemas. As páginas a seguir refletem os resultados dessas discussões. É importante ressaltar que optamos por trazer apenas citações na primeira seção porque ela se relaciona diretamente com as experiências individuais de participantes. Nessa parte específica, buscamos preservar suas percepções de primeira mão sobre os assuntos discutidos, pois acreditamos na importância e na riqueza de suas vozes e contextos para esta pesquisa.

Os dois capítulos seguintes exploram argumentos utilizados para justificar a existência de um sistema de moderação em camadas, bem como propostas para torná-lo uma ferramenta transparente que promova ao mesmo tempo a equidade e a justiça nas plataformas.

Compartilhamento de experiências individuais

Percepções sobre a importância do contexto	"Na América Latina, não há consciência de que não se pode postar qualquer conteúdo porque se trata de uma plataforma privada. Os usuários nem mesmo sabem que existe um mecanismo de apelação em caso de bloqueio. Especialmente no jornalismo, precisamos entender o contexto da linguagem, que pode incluir palavras proibidas pela plataforma, mas usadas em outros contextos."
	"Sou habitante de um país pequeno, e nosso contexto é menos valorizado e considerado na análise da empresa porque moderadores e as políticas não estão envolvidos nem têm conhecimento do contexto."
	"Em 2016, criamos um aplicativo em que pessoas de qualquer cor/etnia podem comprar de produtores negros de várias regiões do Brasil. Ele foi removido porque um professor de direito disse que abriria uma representação no Ministério Público por 'violações de equidade'. Essa postagem se tornou viral, então as plataformas removeram a publicação do aplicativo. Também temos um grupo no Facebook para pessoas negras que discutem questões sociais e políticas. Dentro do grupo, algumas pessoas não refletiram sobre posições políticas, mas se manifestam no grupo porque o consideram um espaço acolhedor e seguro. No entanto, o Facebook percebeu muitas questões discutidas como agressivas. O Facebook tem muita dificuldade em moderar a diversidade, especialmente em uma comunidade que é diversa entre si."
Falta de mecanismos de transparência	"As plataformas também excluíram hashtags usadas no contexto dos protestos na Colômbia e postagens com esse conteúdo. A transparência também é importante. Não recebemos informações factuais das plataformas sobre o motivo das remoções, o que torna difícil questionar a decisão da plataforma."
Capacidade de resposta das plataformas	"Entrar em contato com a plataforma é difícil quando você é um pequeno creator; leva semanas para se obter uma resposta. Às vezes, nem mesmo há uma resposta da plataforma, e o trabalho de creator é prejudicado ao ser desmonetizado sem justificativa."
	"No Instagram, um apresentador de televisão brasileiro com 7 milhões de seguidores disse que a comunidade LGBTQIA+ é desgraçada e que deve ser horrível ter um filho LGBTQIA+ e não poder matá-lo. Esse conteúdo permaneceu na plataforma por muito tempo. Demonstramos que os anunciantes ainda estavam apoiando e ajudando a monetizar aquele conteúdo. Em uma postagem no Instagram e no Facebook, explicamos por que o conteúdo era problemático e fizemos uma reclamação, criticando o conteúdo de ódio que estávamos denunciando. Em questão de minutos, a nossa postagem no Facebook foi removida. Entramos em contato com a Meta e não tivemos sucesso no diálogo. Após a Aliança Nacional LGBTQIA+, parceira da Meta, entrar em contato com a empresa, a Meta restaurou a postagem - mas a campanha já havia perdido engajamento. É importante ressaltar que o conteúdo odioso original denunciado permaneceu na plataforma. Portanto, este é o apelo: ser mais cuidadoso na moderação do conteúdo das denúncias."

Questionamentos sobre o sistema

Quando questionados sobre a necessidade de ter um sistema de moderação em camadas, as/os participantes enfatizaram o fato de que pode ser de interesse público tratar alguns atores de forma diferente com base em critérios específicos. No entanto, para que a estrutura cumpra seu propósito, sua justificativa e padrões devem ser transparentes e públicos. O problema destacado é a falta de transparência, uma vez que o mecanismo por trás do sistema não é divulgado. Essas nuances devem ser ponderadas, porque existem casos em que essa forma de tratamento privilegiado é efetivamente necessária para preservar certas expressões e debates públicos, em oposição a situações em que a moderação prejudicaria a liberdade de expressão dos usuários.

Além disso, as/os participantes dos grupos de discussão mencionaram que estão cientes de que programas que fornecem atenção especial a usuários específicos, frequentemente com base em interesses comerciais, existem em várias plataformas, mas de maneira informal. Isso é visto como problemático porque os métodos empregados não são transparentes e, acima de tudo, geram discriminação, ou seja, a existência de respostas diferentes para situações semelhantes, dependendo dos usuários envolvidos na propagação do discurso.

As sessões também levantaram preocupações em relação aos interesses econômicos das plataformas nas práticas de moderação ao decidir manter ou retirar peças de conteúdo, uma vez que existem certos tipos de expressão, especialmente de parceiros de negócios, que podem impactar sua reputação ou mercados, gerando lucros ou perdas. Quanto dinheiro uma plataforma ganha ao atrasar a moderação de conteúdo inadequado de figuras públicas influentes? Esses valores são importantes para entender se as plataformas estão atrasando intencionalmente bloqueios de conteúdo inapropriado devido ao alto retorno financeiro desse tipo de conteúdo.

Com relação a uma perspectiva regional, as sessões trouxeram considerações sobre a baixa disponibilidade de dados e recursos em alguns países, bem como a falta de difusão e abrangência regional nos relatórios de transparência publicados pelas plataformas. Alguns participantes apontaram que não há dados estruturados suficientes sobre moderação de conteúdo por país ou em outros idiomas.

Por exemplo: quantos usuários cobertos por um sistema de moderação em camadas uma determinada plataforma tem no México? Quantos moderadores por mil usuários? Qual é a diferença no investimento em moderação de conteúdo na Colômbia e na Alemanha? É fundamental ter informações sobre o nível de recursos investidos para analisar a necessidade de sistemas de moderação em camadas e sua extensão. Como seria possível avaliar os impactos de uma tecnologia se não há ferramentas de transparência disponíveis?

Ainda em relação à importância do contexto, as pessoas participantes trouxeram reflexões sobre diferentes aplicações de regras dependendo de regiões específicas. As regras se aplicam a todos os países? Por que os países recebem tratamentos diferentes das plataformas em comparação com outros, por exemplo, ao lidar com desinformação durante períodos eleitorais?

Propostas para o futuro

Ao pensar nos critérios empregados para definir quais tipos de usuários e conteúdos devem contar com um sistema de moderação em camadas, as/os participantes mencionaram a necessidade de a plataforma se comprometer com o mesmo rigor na divulgação e aplicação de suas políticas, independentemente da região, considerando que uma empresa global deve ter capacidade global para fazer cumprir suas regras.

Também é fundamental aplicar esse conjunto de regras com respeito aos contextos e características culturais e locais. As definições que orientam o que pode ou não circular nas plataformas não são universais. Pelo contrário, são culturalmente tendenciosas, baseadas em parâmetros que se aplicam a certas regiões, mas não a outras, o que significa que a remoção de conteúdo pode acabar sendo injustificada em contextos específicos. Um sistema de moderação em camadas bem arquitetado é útil quando leva em consideração nuances regionais.

No âmbito de um sistema de moderação em camadas, a criação de ferramentas, como instâncias de consulta, pode desafiar as dificuldades da relatividade cultural, trazendo mecanismos de verificação, equilíbrio e aprimoramento. Esses espaços poderiam reunir pessoas de populações historicamente marginalizadas, representar públicos locais afetados pelas postagens e promover o estudo da aplicação de regras a contextos específicos.

Além disso, é fundamental que as plataformas divulguem os critérios que motivam a inclusão de determinados conteúdos e usuários em listas de moderação em camadas. A percepção das pessoas participantes é que a seleção de perfis que participam de programas de moderação em camadas não pode se basear exclusivamente na quantidade de seguidores e interesses comerciais das plataformas. As listas devem contemplar, por exemplo, jornalistas, grupos historicamente marginalizados e critérios como alcance do discurso do usuário.

Em conclusão, é fundamental considerar a segurança e a privacidade ao desenvolver ferramentas de transparência em relação a um sistema de moderação em camadas. As pessoas participantes das sessões chamaram a atenção para o fato de que a publicação das próprias listas de usuários selecionados pode ser prejudicial, pois resultaria em

um nível indesejado de exposição, especialmente em casos de pessoas que merecem proteção adicional, como defensores dos direitos humanos e ativistas. Para isso, critérios e dados estatísticos devem ser públicos - gênero, raça, categorias de atores, regiões, entre outros -, mas não os nomes considerados pelo sistema.

O COPO MEIO CHEIO

A pesquisa nos levou a considerar a necessidade de sistemas de moderação em camadas baseados em usuários e/ou conteúdos, a fim de buscar a equidade, em oposição à igualdade formal. É importante tratar indivíduos desiguais de acordo com suas desigualdades. Isso é uma alternativa à moderação em escala e automatizada - que tem potencial para interpretações equivocadas e erros em casos sensíveis - especialmente ao buscar promover os direitos humanos ao proteger discursos políticos e de populações historicamente marginalizadas, jornalismo de interesse público e ativismo.

Além disso, os sistemas de moderação em camadas abrem espaço para pensar nas perspectivas locais. Nos sistemas automatizados de moderação de conteúdo, regras globais são aplicadas independentemente das características culturais e locais. Em outras palavras, os critérios utilizados para manter ou remover conteúdo são concebidos como universais, ignorando realidades sociais, culturais e políticas de outros contextos. Ter diferentes listas e regras para diferentes usuários e conteúdos pode ser útil porque levam em consideração as diferenças, consideram os direitos de populações historicamente marginalizadas e representam públicos locais afetados de maneiras diferentes. Cada contexto tem suas particularidades, e precisamos de regras que levem isso em conta.

Supondo que um país tenha um contexto específico de violação de um certo direito. Defensoras/es desse direito devem desfrutar de maior proteção em seu discurso, especialmente quando representam direitos de populações historicamente marginalizadas, em oposição a países que não têm problemas semelhantes. Existem vários exemplos: considerando a proibição de nudez, o que significa nudez para um país ocidental, em comparação com a perspectiva de um povo indígena brasileiro?

O COPO MEIO VAZIO

Em teoria, a moderação em camadas não deve alterar quais regras são aplicadas, apenas os procedimentos de aplicação. No entanto, na prática, como mostrado pelo caso do *Cross-check*, a aplicação "especial" pode alterar a natureza das decisões sobre o conteúdo, uma vez que acaba implementando resultados diferentes para

alguns indivíduos privilegiados. Assim, pode distorcer uma moderação de conteúdo principiada e consistente em toda a gama de usuários e contextos.

Embora o conceito de implementar um mecanismo como o programa Cross-Check para proteger a pluralidade de discursos em plataformas online seja bem-vindo, sua aplicação pode representar riscos aos direitos humanos e potencialmente proteger práticas comerciais injustas. Por um lado, essa ferramenta pode ser essencial para salvaguardar opiniões e ideias diversas, mas, por outro, pode ser abusada pelas empresas para evitar sua responsabilidade de proteger os direitos humanos. Além disso, as empresas podem usar esses mecanismos para fins de relações públicas, como proteger sua reputação de escândalos de moderação de conteúdo.

Além disso, a pesquisa mostra que há pouca atenção para o impacto da moderação de conteúdo em camadas em nível regional. Em tais contextos, observamos uma falta de literatura e conscientização sobre o uso de sistemas de moderação de conteúdo em camadas para combater a violência contra grupos historicamente marginalizados em diferentes categorias protegidas e marcadores sociais, o que dificulta a realização de conversas construtivas com os *players* do setor, especialmente em regiões como a América Latina. Devido à escassez de dados, faltam estudos que considerem os efeitos do sistema em características políticas, culturais e sociais de países específicos e em diferentes idiomas, por exemplo. Há recursos insuficientes de dados e transparência para algumas regiões em detrimento de outras, e as que são deixadas de lado são justamente aquelas em que os grupos marginalizados lutam mais para acessar um conjunto básico de direitos e garantias. Para concluir, além da falta de transparência, devemos questionar se as plataformas têm incentivo financeiro para atrasar a remoção de conteúdo inadequado. Elas se beneficiam financeiramente desse tipo de atuação? Todos esses fatores devem ser levados em consideração ao avaliar os sistemas de moderação em camadas.

RECOMENDAÇÕES: DAS LISTAS VIP À PROTEÇÃO EQUITATIVA PARA O DISCURSO

Como mencionado, acreditamos que o sistema de verificação em camadas deve existir. Isso se deve à necessidade de uma maior proteção de alguns discursos e figuras, buscando a equidade, e não apenas a igualdade formal. Considerando que a escala é um grande desafio na moderação de conteúdo, e que a tecnologia inevitavelmente será usada para lidar com esse volume, garantir um nível de mecanismo de moderação em camadas para contemplar jornalistas, ativistas e outros atores significa também garantir uma maior proteção de discursos relevantes nas plataformas.

Nesse caso, devemos defender regras e parâmetros mais claros, assim como uma aplicação mais rigorosa em nível mundial. As empresas globais devem ter iniciativa e capacidade de aplicar suas políticas globalmente. Ao pensar em como reformar e melhorar um sistema de moderação em camadas, propomos a inclusão de configurações como:

1. Critérios claros e públicos para estar ou não nas listas de usuários que serão aceitos em programas de moderação em camadas

A operação de um sistema de moderação em camadas deve ser baseada em preceitos de transparência, e a primeira informação importante a ser disponibilizada ao público são os critérios empregados para adicionar ou remover usuários da "lista protegida". O desenvolvimento desses programas e listas não pode ser uma questão de seleção informal que reflita apenas os interesses comerciais das plataformas, por exemplo. Critérios rigorosos devem considerar a proteção do discurso, os perfis dos usuários, o tamanho do mercado e o impacto das publicações, entre outros. Os programas de moderação em camadas não podem ser concebidos como uma permissão para que algumas pessoas tenham mais direitos do que outras.

2. Divulgação das categorias dos perfis e das porcentagens de cada grupo na composição da lista - por exemplo, número de parceiros de negócios, políticos, jornalistas, defensores de direitos humanos, bem como suas regiões, gênero e raça

Além de critérios transparentes, é fundamental que o público tenha acesso a dados agregados sobre as próprias listas, divididos por categorias de perfis, protegendo a identidade dos membros. Esses dados são necessários para uma compreensão mais abrangente do motivo pelo qual determinados tipos de usuários desfrutam de outras camadas de exame. Também ajudam a garantir que esses sistemas não estejam sendo empregados como meras ferramentas de relações públicas ou para fins comerciais. Além disso, a distribuição geográfica de tais programas deve ser divulgada ao público, de modo a promover maior responsabilidade e evitar quaisquer vieses não intencionais que possam surgir da implementação localizada.

3. Transparência em relação ao procedimento e sua lógica, especialmente se há processos de veto de participantes e uma fila de espera para novos participantes, como funciona o processo de entrada e saída, e se é possível inscrever-se ou retirar-se

Existe um procedimento formal que permite que determinados perfis se candidatem a ter camadas adicionais de revisão? Quem decide sobre a inclusão de usuários na lista? É comum que defensores de populações historicamente marginalizadas e de direitos humanos não tenham tantos seguidores quanto as celebridades, por exemplo, mas mereçam padrões mais elevados de proteção de discurso. Essas pessoas teriam a chance de se candidatar a esse grau de proteção, mesmo que seus perfis não sejam tão populares ou comercialmente relevantes quanto outros para as plataformas digitais? As respostas a essas perguntas fornecem legitimidade e se adequam ao direito do usuário de ser informado sobre a mecânica da moderação de conteúdo.

4. Implantação de processos e critérios que levem em conta as particularidades políticas, culturais e sociais de cada região ao adicionar usuários às listas

O fator regional é fundamental para os programas de moderação em camadas, assim como os contextos políticos, culturais e sociais dos usuários. Isso ocorre porque contextos diferentes podem exigir uma aplicação diferente das regras. Por exemplo, se um determinado país tem altos índices de violência contra defensores dos direitos humanos, os critérios devem levar esses números em consideração. Os sistemas em camadas buscam aprimorar o exercício da moderação e, para isso, devem partir das realidades locais para definir suas regras de aplicação.

5. Divulgação periódica de dados sobre a operação dos sistemas, incluindo o número de decisões que foram revertidas pela moderação em camadas, falsos positivos, falsos negativos e assim por diante

A obrigação de divulgar relatórios periódicos de dados sobre os resultados da moderação em camadas é necessária para entender seus impactos e a necessidade de sua existência na operação de plataformas digitais, bem como suas mudanças e evolução ao longo do tempo. A disponibilização dessas informações permitiria que as organizações da sociedade civil, os governos e a academia avaliassem as lacunas da moderação automatizada e criassem ferramentas melhores para corrigir suas falhas.

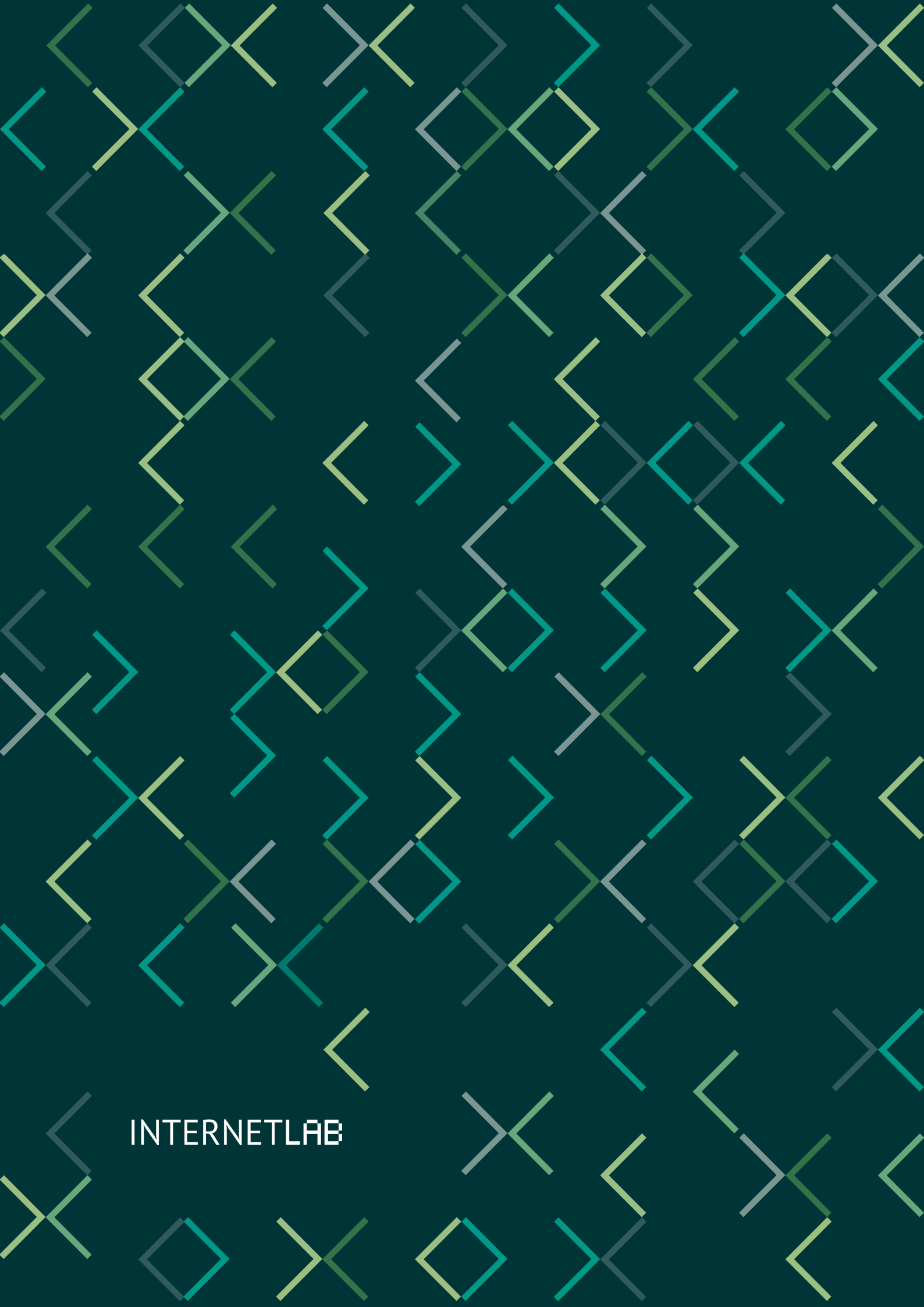
Parte dessas recomendações está alinhada com as emitidas pelo Comitê de Supervisão da Meta no Parecer Consultivo sobre Políticas publicado em dezembro de 2022. Por outro lado, chegamos à conclusão de que é necessário um modelo analítico mais amplo para enquadrar e avaliar plataformas com outros formatos, bem como requisitos específicos de transparência que não foram abordados pelo Comitê.

CONSIDERAÇÕES FINAIS

Nesta opinião consultiva, procuramos desvendar os mecanismos de moderação de conteúdo em camadas, abordando as nuances dos sistemas que regulam a circulação do discurso on-line, bem como a complexidade de tratar usuários de maneiras diferentes. Conforme demonstrado, acreditamos que os sistemas de moderação de conteúdo em camadas devem existir para equilibrar as desvantagens dos sistemas de moderação em escala industrial dentro do complexo exercício logístico de determinar o que deve permanecer e ser removido das plataformas da internet.

Embora esses sistemas possam ser percebidos pela sociedade como problemáticos, pois podem parecer listas VIP que protegem os interesses dos parceiros de negócios das grandes plataformas, é fundamental entender que, ao contrário, quando bem operados, por tratarem de forma diferente usuários diferentes, são capazes de gerar mais equidade e proteção ao discurso.

Com base nesses princípios, formulamos recomendações iniciais de políticas, para que os sistemas de revisão adicionais possam contribuir para promover o acesso à informação e a equidade entre os usuários da plataforma, em vez de causar distorções baseadas em critérios comerciais, que fomentam, ao contrário, a desigualdade no ambiente digital. Os sistemas de moderação em camadas devem prever critérios claros e métricas transparentes, levando em conta os contextos e as realidades locais, evitando que seus objetivos sejam distorcidos para favorecer interesses opacos que poderiam impedir a participação igualitária e o exercício dos direitos humanos online.


INTERNETLAB