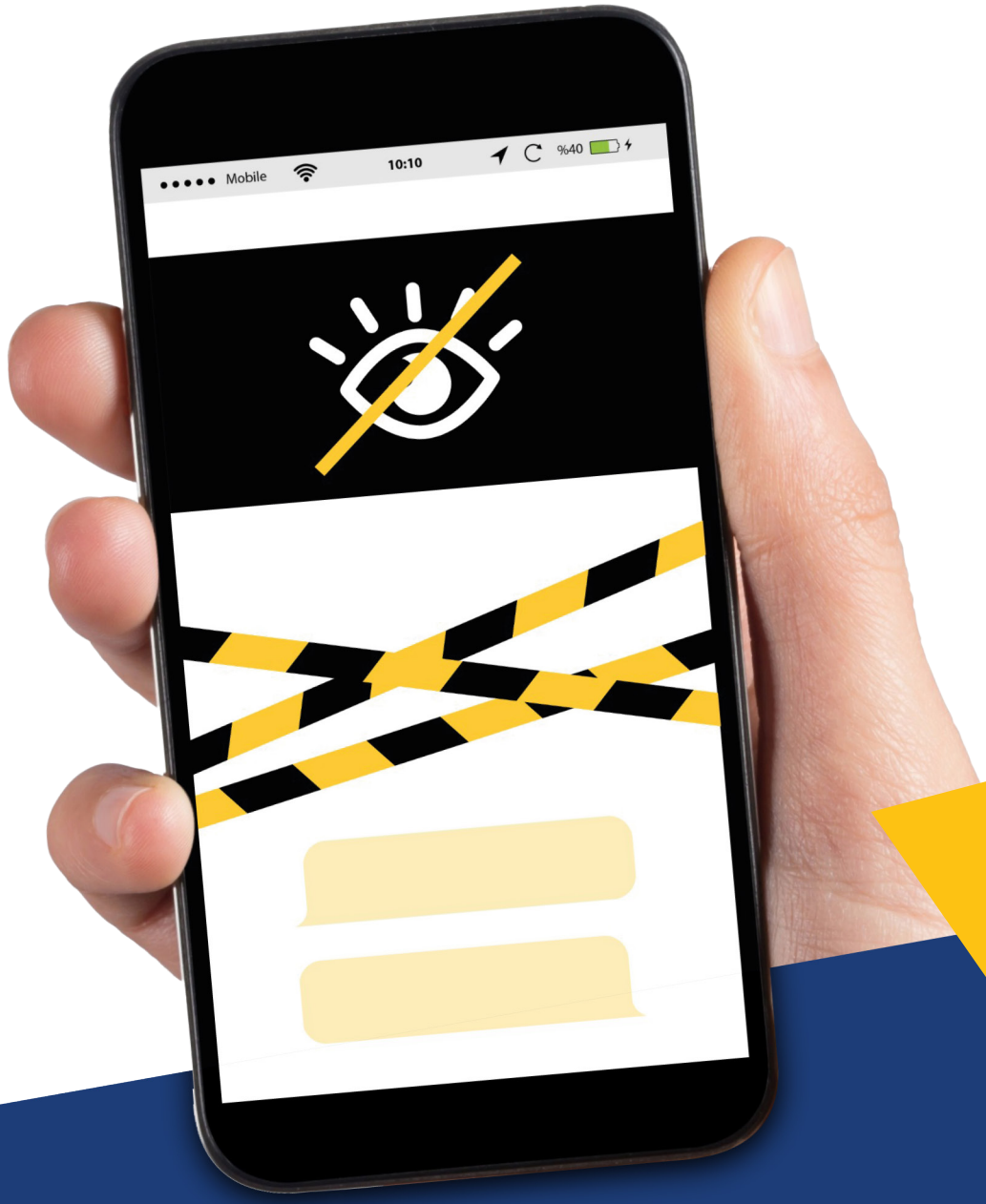

O SHADOW BANNING:
a sutil e oculta censura das grandes plataformas digitais

OBSERVACOM

O shadow banning: a sutil e oculta censura das grandes plataformas digitais

Estudo sobre as práticas de redução de alcance de conteúdos e contas nas redes sociais, e sua relação com a liberdade de expressão online

É uma publicação de
OBSERVACOM
Observatorio Latinoamericano de Regulación, Medios y Convergencia
Av. Libertador 1878 apto. 715
Montevideo, Uruguay

www.observacom.org

Com o apoio de

e de

Carolina Martínez Elebi - Autora

É Licenciada em Ciências da Comunicação pela Universidade de Buenos Aires (UBA) da Argentina, onde atua como docente desde 2011. É consultora em temas vinculados ao impacto das tecnologias da informação e comunicação nos direitos humanos e é diretora do meio digital DHyTecno. Em 2018, cursou o Programa de Direito da Internet e Tecnologias das Comunicações do Centro de Estudos em Tecnologia e Sociedade (CETyS-UdeSA). É coordenadora acadêmica da Pós-graduação em Inteligência Artificial e Sociedade da Universidade Nacional de Três de Fevereiro (UNTREF) e integra o Observatório de Impactos Sociais da Inteligência Artificial dessa mesma universidade.

Vladimir Cortés Roshdestvensky - Autor

É especialista em direitos humanos com mais de uma década de experiência em direitos digitais, liberdade de expressão e governança democrática. Com um Mestrado em Direitos Humanos pela Universidade de Pádua, liderou pesquisas sobre IA, exclusão digital e moderação de conteúdo. Atualmente é Diretor de Campanhas e Alianças na Digital Action, onde coordena estratégias para exigir prestação de contas de governos e empresas de tecnologia. Recebeu a bolsa de Líderes do LACNIC e “Nicola Tonon”. Trabalhou na ARTICLE 19 e como analista para o relatório Freedom on the Net da Freedom House.

01.

INTRODUÇÃO

No atual ecossistema digital, as plataformas exercem um papel central como intermediárias na circulação de informações. Como parte dessa função, elas implementam sistemas de moderação de conteúdo que incluem medidas visíveis e relativamente conhecidas, como a remoção de publicações, a suspensão temporária ou definitiva de contas e outras sanções que, em geral, são notificadas às pessoas usuárias. Essas decisões costumam vir acompanhadas de mecanismos de apelação, pelo menos nos termos estabelecidos pelas próprias empresas, e se inserem no cumprimento de suas normas comunitárias.

No entanto, nos últimos anos, essas formas tradicionais de moderação têm sido complementadas — e, em alguns casos, substituídas — por práticas igualmente impactantes, embora mais sutis, menos transparentes e muito mais difíceis de identificar. Já não se trata apenas da remoção direta, mas de intervenções sutis e pouco claras que afetam a circulação de informações de interesse público e outras formas de conteúdo gerado pelas pessoas usuárias. A Relatoria Especial para a Liberdade de Expressão da Comissão Interamericana de Direitos Humanos (RELE) advertiu que as empresas de tecnologia devem evitar que algoritmos e sistemas automatizados — especialmente aqueles que operam sem supervisão humana significativa — se tornem uma ameaça à liberdade de expressão. O risco é particularmente grave quando suas decisões impõem restrições excessivas e desproporcionais, que afetam com maior frequência grupos historicamente marginalizados.[1]

O *shadow banning* representa esse tipo de tática. Embora mantenham tecnicamente o conteúdo disponível, essas práticas reduzem drasticamente sua visibilidade e afetam meios de comunicação, ativistas, empreendedores e pessoas usuárias que, em muitos casos, nem sequer têm consciência dessa limitação. O que preocupa é que essa redução de visibilidade funciona como uma forma de censura silenciosa. Sem notificações nem explicações claras, vozes diversas acabam efetivamente excluídas do debate público digital, o que compromete o pluralismo informativo e o debate democrático.

O *shadow banning*, embora não bloqueie completamente a capacidade de expressão dos usuários, tem o potencial de afetar significativamente quatro dimensões cruciais do discurso: a disponibilidade do conteúdo, sua visibilidade no ecossistema digital, sua acessibilidade para diferentes públicos e, em última instância, sua capacidade de gerar impacto no debate público.

Este alerta ganha relevância especial à luz do critério estabelecido pela Corte Interamericana de Direitos Humanos (CIDH), que consagrou um princípio fundamental: a expressão e sua difusão constituem um todo indivisível. Essa interpretação amplia significativamente o alcance protetivo do direito à liberdade de expressão, garantindo não apenas a possibilidade de manifestar ideias e opiniões, mas também o direito essencial de utilizar qualquer meio apropriado para que essas ideias alcancem o maior número possível de

1 Comissão Interamericana de Direitos Humanos (CIDH), Relatoria Especial para a Liberdade de Expressão. (2024). *Inclusión digital y gobernanza de contenidos en internet.* Organización de los Estados Americanos. p. 69, párr. 276. https://www.oas.org/es/cidh/expresion/informes/Inclusion_digital_esp.pdf

pessoas — e para que esses potenciais receptores possam, de fato, ter acesso a essa informação.[2]

Essa transformação na moderação levanta sérias questões sobre transparência e prestação de contas. Quando as plataformas reduzem algoritmicamente o alcance de certos conteúdos sem notificação, quem supervisiona essas decisões? Com base em quais critérios elas são implementadas?

Este estudo busca desvendar o impacto dessas práticas sobre meios de comunicação, jornalistas, ativistas e usuários, explorando também até que ponto as plataformas são transparentes em relação a esses mecanismos que, embora invisíveis, reconfiguram profundamente nosso ecossistema de informação e impactam o estado psicoemocional das pessoas.

O objetivo desta pesquisa é examinar o fenômeno do *shadow banning* em plataformas digitais, por meio da identificação de casos específicos e da análise de seu impacto na visibilidade de meios de comunicação, vozes críticas e setores sub-representados. Além disso, a pesquisa busca avaliar o grau de transparência das plataformas em relação a essas práticas e sua incidência na participação democrática nos espaços públicos digitais.

Esta investigação, focada em identificar as práticas de *shadow banning* na América Latina, adotou um método predominantemente qualitativo, baseado em entrevistas semiestruturadas com diversos atores do ecossistema digital, na análise dos termos e condições das grandes empresas de tecnologia como a Meta (Instagram e Facebook) e a X (antigo Twitter), bem como numa revisão bibliográfica sobre a temática. A análise documental das políticas e termos de serviço permitiu confrontar as experiências relatadas com o marco normativo declarado pelas próprias plataformas, e a partir disso identificamos discrepâncias significativas entre as práticas percebidas pelas pessoas usuárias e as políticas explícitas das empresas. Adicionalmente, a revisão bibliográfica abrangeu tanto literatura acadêmica quanto documentos de trabalho de organizações da sociedade civil e *think tanks*, o que nos permitiu contextualizar o fenômeno dentro do debate global sobre moderação de conteúdo e liberdade de expressão em ambientes digitais.

Esse método nos proporcionou a oportunidade de documentar e validar as experiências em primeira mão de ativistas, jornalistas, defensores de direitos humanos e empreendedores digitais que relataram terem sido submetidos a restrições de visibilidade pouco transparentes em diversas plataformas. Por meio desses depoimentos, conseguimos identificar padrões comuns nas experiências relatadas e reconhecer impactos concretos no exercício da liberdade de expressão em ambientes digitais.

É importante destacar que a investigação do *shadow banning* enfrenta desafios metodológicos consideráveis, tanto do ponto de vista técnico quanto da coleta de dados. Atualmente, as principais plataformas digitais impõem restrições substanciais ao acesso para

2 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet*, párr. 276.

fins de pesquisa, seja por meio de processos complexos para solicitação de dados, da limitação no uso de APIs, do encerramento de ferramentas como o CrowdTangle — que anteriormente facilitavam a pesquisa independente — ou por uma redução generalizada na transparência de suas operações algorítmicas. Essas barreiras técnicas dificultam enormemente a possibilidade de documentar o fenômeno do *shadow banning* com evidências quantitativas robustas, o que nos levou a priorizar a documentação qualitativa de casos representativos.

A natureza inerentemente opaca do *shadow banning* constitui, talvez, o maior obstáculo metodológico. Diferentemente de outras formas de moderação de conteúdo, nas quais há uma notificação explícita, o *shadow banning* se caracteriza justamente pela ausência de transparência e de comunicação com as pessoas usuárias afetadas. Essa característica, somada à ambiguidade na redação das políticas das plataformas — que raramente mencionam essas práticas de forma explícita, ou utilizam eufemismos como "redução de distribuição" ou "ajustes de visibilidade" —, cria um cenário em que a documentação sistemática se torna extraordinariamente complexa. Por isso, nossa metodologia se concentrou em triangular as experiências relatadas com mudanças observáveis no alcance e na visibilidade do conteúdo, reconhecendo as limitações inerentes à investigação de práticas deliberadamente projetadas para serem imperceptíveis. Vale destacar que a literatura acadêmica e técnica sobre *shadow banning* está disponível majoritariamente em inglês, sendo ainda um campo relativamente pouco explorado em espanhol.

02.

O SHADOW BANNING: A INVISIBILIZAÇÃO OCULTA DE CONTEÚDOS E USUÁRIOS

No espaço digital dominado pelas grandes plataformas tecnológicas — onde milhões de pessoas conversam, compartilham e acessam informações — opera uma forma silenciosa de silenciamento que afeta ativistas, jornalistas e outros usuários, embora poucos consigam detectá-la a tempo. Trata-se do *shadow banning*, um conjunto de práticas de moderação por meio das quais as plataformas reduzem discretamente o alcance e a visibilidade de determinados perfis ou publicações, sem notificar a pessoa afetada. Por isso, se fala de uma forma de moderação "oculta" em dois sentidos: ela invisibiliza o conteúdo para os demais usuários e também oculta a sanção de quem a sofre. Como um fantasma que percorre os algoritmos, o *shadow banning* deixa suas vítimas presas em um limbo comunicacional: continuam falando, mas sua presença se desvanece sem deixar rastro.

Diferentemente das restrições explícitas — como a exclusão de conteúdos ou a suspensão de contas —, o *shadow banning* mantém uma aparência de normalidade. As contas afetadas podem continuar publicando normalmente, mas seus conteúdos são progressivamente excluídos das conversas públicas: desaparecem dos resultados de busca, tornam-se invisíveis nas hashtags mais populares ou deixam de aparecer no *feed* dos próprios seguidores.[3] Trata-se, em muitos aspectos, de uma forma mais sofisticada de silenciamento digital.

As manifestações do *shadow banning* são tão diversas quanto sutis: desde uma queda repentina nas interações até o desaparecimento completo dos sistemas de recomendação ou dos resultados de busca. Um jornalista que normalmente recebe centenas de comentários pode se ver, de um dia para o outro, falando para o vazio; uma ativista que utiliza *hashtags* relacionadas aos direitos humanos descobre que suas publicações nunca aparecem nessas buscas; uma educadora sexual vê seus conteúdos informativos sendo filtrados por algoritmos que os classificam como "conteúdo limítrofe".

O mais preocupante é que essas restrições algorítmicas não são aplicadas de forma equitativa. As evidências indicam que elas afetam desproporcionalmente comunidades marginalizadas: ativistas sociais e políticos, jornalistas independentes, pessoas LGBTQIA+ e educadores em sexualidade integral. O problema central é a opacidade. Sem notificações, explicações ou mecanismos eficazes de apelação, as pessoas afetadas são forçadas a realizar um trabalho invisível e exaustivo: desde formular hipóteses sobre o funcionamento do algoritmo até modificar sua linguagem (*algospeak*) para evitar punições, ou construir redes coletivas para verificar e driblar a censura encoberta.

Esse fenômeno não é uma raridade técnica, mas uma ameaça concreta aos direitos fundamentais à liberdade de expressão — especialmente em contextos como o da América Latina, onde a visibilidade digital pode ser crucial para o ativismo, a denúncia ou o exercício democrático. É urgente, portanto, avançar na construção de mecanismos mais robustos de prestação de contas e

3 Nicholas, G. (2022). *Shedding light on shadow banning*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2022/04/remediated-final-shadowbanning-final-050322-upd-ref.pdf

transparência sobre as decisões tomadas por aqueles que hoje controlam as portas de acesso à informação pública.

As manifestações mais comuns de *shadow banning* incluem:

- **Redução do alcance das publicações da pessoa usuária:** Os conteúdos publicados passam a alcançar um público significativamente menor do que o habitual, mesmo sem terem sido excluídos. Por exemplo, uma publicação que normalmente receberia centenas ou milhares de interações (curtidas, comentários ou compartilhamentos) passa a ter muito menos, porque a plataforma decide não mostrá-la nos *feeds* de outros usuários ou não a prioriza o suficiente para que seja vista. Ela é posicionada mais abaixo — ou no final —, o que, na prática, significa que quase ninguém a verá. Essa redução pode ocorrer de forma repentina ou gradual nas curtidas, comentários e visualizações, de maneira desproporcional ao número de seguidores e às métricas típicas de engajamento. Alguns autores documentaram que "as publicações daqueles que relataram estar [bloqueados na sombra] não aparecem nos feeds de seus seguidores de forma alguma, e aparentemente são completamente desprezadas pelo algoritmo".[4]

- **Restrição da visibilidade da pessoa usuária no resultado da busca:** Mesmo que um usuário digite exatamente o nome da conta na ferramenta de busca, ela não aparece nos resultados, não é sugerida[5] ou aparece muito abaixo, o que dificulta sua localização. Isso limita o crescimento orgânico da conta e sua participação nos debates públicos.

- **Remoção da conta das sugestões e recomendações visíveis para outros usuários:** a plataforma deixa de incluir essa conta em seções como "Pessoas que talvez você conheça", "Sugestões para você", "Contas recomendadas" ou similares. Por exemplo, uma conta deixa de ser sugerida a usuários que poderiam se interessar por seu conteúdo, o que reduz sua chance de alcançar novas audiências.

- **Exclusão de *hashtags*, *feeds* de descoberta e tendências:** Mesmo que a pessoa marque sua publicação com uma *hashtag* específica, ela não aparece quando outros usuários clicam nessa *hashtag*. Por exemplo, um usuário que publica com as hashtags #FreePalestine ou #MeToo pode perceber que seu post não aparece entre os resultados dessa etiqueta, o que impede que sua mensagem se some à conversa pública. Isso também se refere à ausência das publicações em páginas de descoberta algorítmica, como a aba "Explorar" do Instagram ou a página "Para você" do TikTok.

- **Limitação na interação com outros usuários:** Os comentários ou respostas do usuário aparecem ocultos ou rebaixados para os demais, embora possam ser visíveis para o próprio autor. Por exemplo, um jornalista

4 Blunt, D., Wolf, A., Coombes, E., e Mullin, S. (2020). *Posting into the void: studying the impact of shadowbanning on sex workers and activists.* Hacking//Hustling.

5 Le Merrer, E., Morgan, B. e Trédan, G. (2021). Setting the record straighter on shadow banning. En *IEEE Infocom 2021-IEEE Conference on Computer Communications* (pp. 1-10). IEEE. https://arxiv.org/pdf/2012.05101

comenta uma publicação viral, mas seu comentário é invisível para o resto das pessoas, reduzindo sua capacidade de participar de conversas públicas.

- **Bloqueio de funcionalidades:** Impossibilidade de utilizar certas ferramentas que permitem a interação com outros usuários. Por exemplo, não conseguir curtir, responder a publicações de outras pessoas, ou que uma publicação não esteja vinculada ao nome da conta que a fez.[6]

Todas essas ações não apenas fazem com que os conteúdos tenham uma circulação ou alcance reduzido, como também diminuem ou impedem a possibilidade de serem descobertos — tanto das publicações quanto da própria conta e presença na rede, dificultando o crescimento da audiência ou da comunidade de seguidores.

O problema dessa prática, implementada pelas plataformas como forma de punir o suposto descumprimento das suas normas comunitárias, é que ela limita a circulação das expressões das pessoas usuárias sem que elas percebam. Diferentemente de um bloqueio explícito — no qual a conta é suspensa ou excluída de forma direta e, geralmente, com uma notificação da plataforma —, essas restrições operam de maneira silenciosa. Embora a conta não esteja bloqueada em termos formais, ao restringir a circulação de suas expressões e limitar sua capacidade de ser descoberto por novas audiências, o impacto é o mesmo: sua participação no debate público online é dificultada — ou diretamente excluída.

O *shadow banning* é aplicado de forma automatizada por algoritmos baseados em inteligência artificial que "moderam" a circulação de conteúdos para controlar o discurso dentro da plataforma. Sua falta de transparência e a ausência de mecanismos claros para detectá-lo ou contestar suas decisões fazem com que ele seja uma das formas mais polêmicas de intervenção dessas empresas nos novos espaços públicos da internet.

Essas práticas afetam gravemente a liberdade de expressão e o pluralismo informativo, já que meios de comunicação, jornalistas e ativistas podem ter seu impacto na conversa pública reduzido, sem contar com mecanismos adequados e oportunos para reivindicar seus direitos e reverter as medidas.

Para compreender adequadamente o fenômeno do *shadow banning*, é fundamental distinguir dois conceitos-chave que operam nas plataformas digitais como X, Instagram, Facebook, entre outras: a moderação de conteúdo e a curadoria de conteúdo.

O primeiro refere-se ao conjunto de políticas, sistemas e ferramentas que as plataformas implementam para gerenciar o conteúdo gerado por seus usuários, determinando o que é publicado, o que é removido e como esse conteúdo é controlado.[7] Esse processo pode ser

6 Blunt, D., Wolf, A., Coombes, E. e Mullin, S. (2020). *Posting into the void: studying the impact of shadowbanning on sex workers and activists.*

7 Center for Democracy & Technology. (2021). *Outside looking in: approaches to content moderation in end-to-end encrypted systems.* https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/

estruturado de três formas: 1) centralizada (como no X, Facebook ou YouTube), em que a própria plataforma aplica internamente as regras; 2) "distribuída" (como no Reddit ou na Wikipedia), em que as próprias comunidades gerenciam a moderação com mínima intervenção da plataforma; ou 3) "híbrida" (como no Twitch), que integra ambos os enfoques. A moderação, de forma geral, desenvolve-se por meio de fases sequenciais que vão desde a definição de regras até os mecanismos de apelação, e pode ser aplicada tanto antes da publicação (*ex ante*) quanto depois (*ex post*).[8]

Para definir o conceito de 'moderação de conteúdo em redes sociais', a Relatoria Especial para a Liberdade de Expressão da Comissão Interamericana de Direitos Humanos (RELE) adota definições provenientes do processo de Diálogo das Américas, assim como de documentos elaborados por organizações da sociedade civil especializadas no tema. Assim, no parágrafo 187 do relatório *Inclusão digital e governança de conteúdo na Internet*, publicado em junho de 2024,[9] "moderação de conteúdo" é definida como a prática organizada de filtrar o conteúdo gerado e visualizado pelos usuários e publicado em plataformas digitais. O relatório enumera diferentes tipos de moderação de conteúdo: pré-moderação, pós-moderação, moderação reativa, moderação distribuída e moderação automatizada.

O relator também destaca em seu relatório que "o processo de moderação pode ser realizado por meio da ação direta de uma pessoa ou por processos automatizados baseados em ferramentas de inteligência artificial, juntamente com o processamento de grandes quantidades de dados das pessoas usuárias".[10] A moderação pode implicar "a remoção do conteúdo de forma permanente ou temporária, em toda a plataforma ou em relação a determinados grupos de usuários em uma área geográfica específica, ou ainda afetar contas de usuários sob diferentes modalidades". Outro tipo de moderação pode envolver ações como rotular conteúdos, fornecer informações adicionais e contextualizadas sobre uma publicação, ou desmonetizar os conteúdos, entre outras medidas.

A curadoria de conteúdo, por outro lado, é o processo pelo qual as plataformas digitais selecionam, organizam e apresentam conteúdos a uma audiência, de acordo com critérios que são desconhecidos pelos usuários. Ela determina quais conteúdos terão maior visibilidade e quais serão relegados nos *feeds*, nos resultados de busca e nas recomendações personalizadas das pessoas usuárias da plataforma.

A RELE considera a curadoria de conteúdo "como decisões automatizadas sobre o alcance, a classificação, a promoção ou a visibilidade dos conteúdos. As plataformas geralmente curam os conteúdos com base em recomendações personalizadas para os perfis das pessoas usuárias".[11] Ao mesmo tempo,

8 Klonick, K. (2018). *The new governors: the people, rules, and processes governing online speech.* St. John's University School of Law. https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1184&context=faculty_publications
9 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet.*
10 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet*, párr. 187.
11 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet*, párr. 188.

adverte: “Na medida em que certos conteúdos são privilegiados, a curadoria pode acabar por amplificar ou reduzir o alcance de determinados discursos”.[12]

Nesse sentido, a curadoria de conteúdo não é neutra, pois responde a critérios definidos por cada plataforma, influenciando o que os usuários podem ver — e o que permanece oculto.

Esses processos são, em sua maioria, automatizados e gerenciados por sistemas algorítmicos e de inteligência artificial que analisam a atividade dos usuários para decidir quais conteúdos promover e quais limitar, com base em critérios como “fazer com que seja um lugar seguro para inspiração e expressão”. No entanto, esses critérios respondem ao modelo de negócios e aos interesses comerciais das *big techs*, voltados a capturar a atenção dos usuários e mantê-los o maior tempo possível nas plataformas. Mais recentemente, mudanças nas políticas de várias das principais plataformas digitais confirmaram que também existem decisões políticas envolvidas na definição desses critérios.

O *shadow banning* ocupa uma posição singular no espectro das práticas de governança de conteúdo, situando-se na interseção entre a moderação e a curadoria de conteúdo. Ele não constitui uma exclusão completa do conteúdo (moderação tradicional), mas sim uma intervenção algorítmica que reduz significativamente sua visibilidade ou alcance (curadoria negativa).

Especificamente, o *shadow banning* se insere predominantemente no campo da curadoria de conteúdo, pois afeta diretamente a forma como o material é distribuído e apresentado a outros usuários, sem que ele seja removido. No entanto, quando essa redução de visibilidade é aplicada como consequência de supostas infrações às normas da comunidade, ela também funciona como uma forma de moderação *ex post*, menos severa do que a exclusão completa. A característica definidora do *shadow banning* — e o que o torna particularmente problemático do ponto de vista dos direitos humanos — é sua natureza deliberadamente opaca: diferentemente de outras medidas de moderação, nas quais a plataforma notifica o usuário sobre a ação tomada, o *shadow banning* opera intencionalmente sem transparência, deixando a pessoa usuária na incerteza sobre por que seu conteúdo não está alcançando seu público habitual.

12 CIDH. (2024). Inclusión digital y gobernanza de contenidos en internet, párr. 188.

03.

ESTUDOS E EVOLUÇÃO CONCEITUAL DO SHADOW BANNING

O *shadow banning* passou por uma evolução conceitual significativa desde suas origens nos primeiros fóruns da internet. Inicialmente, o termo se referia especificamente a uma técnica de moderação pela qual os comentários e publicações de usuários considerados "problemáticos" — ou seja, identificados como assediadores ou *trolls* — eram ocultados para o restante da comunidade, enquanto, para a pessoa afetada, mantinha-se a ilusão de que seu conteúdo continuava visível.[13] Essa estratégia tinha como objetivo principal evitar que as pessoas sancionadas criassem novas contas ao perceberem a restrição.

Savolainen oferece uma abordagem analítica mais sociocultural ao tentar compreender o *shadow banning* como um elemento do "folclore algorítmico", ou seja, um conjunto de "crenças e narrativas sobre os sistemas de moderação que são transmitidas informalmente e podem existir em tensão com os relatos oficiais".[14] Essa perspectiva destaca como o termo funciona como um ponto de articulação discursiva para experiências diversas - porém conectadas - de governança de plataformas, unificadas por uma sensação compartilhada de opacidade e incerteza.

Outras pesquisas têm ressaltado como esse tipo de restrição afeta desproporcionalmente usuários de comunidades marginalizadas. Em particular, alguns autores sugerem que os usuários desenvolveram o que chamam de "teorias algorítmicas populares"[15] para tentar decifrar como funcionam os algoritmos das plataformas. Essas estratégias incluem mudanças no uso de hashtags, alterações em imagens ou até mesmo a criação de contas secundárias para verificar se foram alvo de *shadow banning*.

A pesquisa de Kojah e outros autores, por sua vez, aprofunda essa linha ao definir o *shadow banning* como "uma forma controversa de governança de plataformas caracterizada pelo uso de algoritmos opacos para reduzir ou degradar conteúdo".[16] Sua análise caracteriza essa prática como uma forma de censura "insidiosa, porém sutil", que impacta múltiplas dimensões da experiência dos usuários — desde a visibilidade e os ganhos financeiros até a saúde mental e as comunicações interpessoais.

13 Cole, S. (31 de julho de 2018). *Where Did the Concept of «Shadow Banning» Come From?* VICE. https://www.vice.com/en/article/where-did-shadow-banning-come-from-trump-republicans-shadowbanned/

14 Savolainen, L. (2022). Algorithmic lore and the myths of non-promotion. *Information, Communication & Society*, 25(8), p. 1096.

15 Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., & Haimson, O. L. (2024). "What are you doing, TikTok?": How marginalized social media users perceive, theorize, and "prove" shadowbanning. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), Article 154. https://doi.org/10.1145/3637431

16 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. e Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 9(1), 1-22. p. 19. https://hdl.handle.net/2027.42/196532

Características transversais do *shadow banning*

Independentemente da manifestação específica, os estudos coincidem em identificar certas características definidoras do *shadow banning*:

- **Opacidade**: Ausência de notificação ou explicação por parte da plataforma sobre a restrição.
- **Gradação variável**: O *shadow banning* não é binário; ele existe em um espectro de redução de visibilidade.
- **Acúmulo de efeitos: Diferentes tipos de** *shadow ban* podem coexistir, ampliando seu impacto.
- **Detecção indireta:** Usuários desenvolvem métodos informais para verificar se estão sendo *shadow banned*, como comparar sua visibilidade com a de outras pessoas ou usar ferramentas de terceiros.

A tipologia descrita anteriormente, somada a essas características transversais, reflete a complexidade e a evolução das práticas de moderação algorítmica nas plataformas digitais. E mostra como a diversificação do termo *shadow banning*, para além de sua definição original, reflete a forma como as pessoas usuárias conceituam e respondem a formas emergentes de governança algorítmica marcadas pela opacidade e pela incerteza.[17]

Impactos desiguais: grupos afetados e dinâmicas de marginalização

Uma constatação consistente na literatura revisada é que o *shadow banning* afeta desproporcionalmente grupos já marginalizados. Algumas pesquisas indicam que esse tipo de moderação é mais comum em conteúdos relacionados à sexualidade, identidade racial ou protestos sociais. Por exemplo, o Instagram já foi criticado por censurar imagens de corpos femininos, incluindo publicações de ativistas do movimento Free the Nipple.[18]

Algumas investigações documentaram especificamente como as políticas de conteúdo "no limite" (*borderline content*, em inglês) impactam negativamente comunidades vulneráveis, como trabalhadoras sexuais, educadoras sexuais e membros da comunidade LGBTQIA+. Reportagens jornalísticas sugeriram padrões semelhantes em relação a pessoas negras,[19] mulheres[20] e comunidades Queer.[21]

17 Savolainen, L. (2022). Algorithmic Lore and the Myths of Non-Promotion. Information.
18 Are, C. (2021). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22(8), 2002-2019. p. 2002.
19 BBC. (2020). *Facebook and instagram to examine racist algorithms.* British Broadcasting Corporation. https://www.bbc.com/news/technology-53498685
20 Cook, J. (2019). *Instagram's shadow ban on vaguely 'inappropriate' content is plainly sexist.* Huffington Post. https://www.huffpost.com/entry/instagram-shadowbansexist_%20n_5cc72935e4b0537911491a4f
21 Joseph, C. (2019). Instagram's murky 'shadow bans' just serve to censor marginalised communities.

Um estudo baseado em diários e entrevistas com oito criadores de conteúdo em situação de marginalização documentou que "criadores com identidades marginalizadas (mulheres, dançarinos de pole dance, pessoas gordas ou LGBTQIA+) são desproporcionalmente afetados pelo *shadow banning*".[22] Os participantes percebiam que "pessoas com identidades marginalizadas eram mais afetadas pelo *shadow banning* e por outras formas de censura do que homens que criavam conteúdos semelhantes, e do que pessoas que se encaixam nos padrões convencionais de beleza".[23]

A pesquisa de Are oferece uma análise particularmente detalhada sobre como o *shadow banning* afeta dançarinas de pole dance no Instagram, revelando como a censura de conteúdo relacionado ao pole dance reflete percepções enviesadas sobre o corpo feminino e as expressões de sexualidade. Seu estudo destaca como até mesmo atividades artísticas e esportivas podem ser erroneamente categorizadas como "conteúdo sexualmente sugestivo"[24] quando realizadas por certos corpos ou por determinadas comunidades.

O trabalho invisível da pessoa usuária sob *shadow banning*

Uma contribuição particularmente valiosa da literatura recente é a identificação do trabalho invisível que as pessoas usuárias precisam realizar para navegar, mitigar e se adaptar aos sistemas opacos de moderação. A pesquisa oferece um aporte significativo ao identificar três categorias específicas desse trabalho invisível:

1. **Trabalho mental e emocional:** A carga cognitiva e psicológica de estar constantemente antecipando que tipo de conteúdo pode ser restringido. Como explica um participante do estudo: "Desviar minha atenção da criação de conteúdo para resolver problemas de alinhamento com as políticas funciona como uma distração, o que prejudica minha consistência na produção".[25]

2. **Trabalho sem rumo:** Esforços feitos na esperança de evitar **o** *shadow banning*, mas que não contribuem diretamente para a criação de conteúdo criativo. Isso inclui práticas como postar *selfies* "para agradar o algoritmo" após conteúdos potencialmente controversos ou com alto risco de serem restringidos — como publicações pró-Palestina —, ou ainda usar

The Guardian. https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive

22 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. e Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 2.

23 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. e Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 9.

24 Are, C. (2021). The Shadowban Cycle. p. 2004.

25 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. e Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p.13.

algospeak[26] (modificar palavras potencialmente problemáticas).

3. **Trabalho comunitário:** O esforço colaborativo entre criadores para compartilhar estratégias e se apoiarem mutuamente. Como descreve um participante do estudo: "Tenho trabalhado com outros criadores para ajudar a impulsionar o engajamento durante um *shadow ban*. Promovemos e interagimos com os conteúdos uns dos outros durante uma suspensão suspeita."[27]

26 Este termo, segundo Kojah et al. (2025), refere-se a "escrever palavras de forma maliciosa ou bloquear imagens para enganar o algoritmo e contornar a moderação e supressão de conteúdos, uma prática que requer muito tempo e esforço". (p. 14).

27 Kojah, S. A., Zhang, B. Z., Are, C., Delmonaco, D. e Haimson, O. L. (2025). «Dialing it back»: Shadowbanning, invisible digital labor, and how marginalized content creators attempt to mitigate the impacts of opaque platform governance. p. 16.

04.

IMPACTO NOS MEIOS DE COMUNICAÇÃO E NAS PESSOAS USUÁRIAS

O *shadow banning* opera como uma forma de censura digital particularmente perniciosa: invisível, pouco documentada e devastadora para quem relata vivenciá-la. Na América Latina, ativistas, educadores sexuais, jornalistas e pequenos empreendedores experimentam uma queda drástica no alcance de suas publicações — que antes chegavam a milhares de visualizações e, muitas vezes, tratavam de temas de interesse público — e que, de repente, passam a alcançar apenas algumas centenas de pessoas, sem qualquer explicação ou aviso prévio. Essa redução algorítmica de alcance não apenas compromete a visibilidade do conteúdo, mas também provoca profundos impactos psicoemocionais, marcados por ansiedade constante, sentimentos de impotência e uma autocensura preventiva que silencia vozes importantes no debate público.

O criador afetado fica preso em um limbo digital no qual continua publicando para uma audiência fantasma, investindo tempo e recursos em um esforço que as plataformas silenciosamente condenaram à irrelevância.

A questão se torna ainda mais grave quando percebemos que a moderação realizada pelas grandes empresas de tecnologia afeta temas como saúde sexual, diversidade corporal, cobertura jornalística ou crítica política, justamente os discursos que usufruem de proteção especial[28] no sistema interamericano de direitos humanos.

Além disso, essa incerteza constante obriga as pessoas usuárias a desenvolver diferentes estratégias — às vezes sozinhas, outras vezes de forma coletiva —, bem como a investir recursos não para melhorar seus conteúdos ou produtos, mas para seguir em uma luta desigual contra sistemas algorítmicos opacos que condenam sua presença digital a um canto obscuro.

O que poderia ser considerado uma simples decisão empresarial sobre a distribuição de conteúdos transforma-se em um poderoso mecanismo de controle social. Somado à impossibilidade de prever o que desencadeará restrições — desde mencionar termos como *maconha, educação sexual, Free Palestine*, até mostrar corpos não normativos —, isso gera um efeito inibitório, no qual as pessoas passam a considerar

28 A jurisprudência do sistema interamericano de direitos humanos estabeleceu três categorias de discursos especialmente protegidos, reconhecendo seu papel fundamental no fortalecimento da democracia e no exercício pleno dos direitos humanos. Em primeiro lugar, a proteção ao discurso político e sobre temas de interesse público, considerando que esse tipo de expressão é essencial para a formação de uma opinião pública informada e para a participação das pessoas nos processos democráticos e nos assuntos públicos. Em segundo lugar, a proteção reforçada ao discurso sobre agentes públicos no exercício de suas funções e sobre pessoas candidatas a cargos públicos, entendendo que o escrutínio das ações daqueles que ocupam ou aspiram a posições de poder é crucial para a transparência e a prestação de contas. Por fim, a proteção especial ao discurso que constitui um elemento da identidade ou dignidade pessoal de quem se expressa, reconhecendo assim a importância da liberdade de expressão para o desenvolvimento individual e a autonomia pessoal. Essa proteção diferenciada busca garantir que esses tipos de discurso — fundamentais para o debate público e para a realização pessoal — não sejam indevidamente restringidos, promovendo, assim, uma sociedade mais aberta, plural e democrática. Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos [CIDH]. (2009). *Marco jurídico interamericano del derecho a la libertad de expresión* (párr. 32-56). Organización de los Estados Americanos. https://www.oas.org/es/cidh/expresion/docs/publicaciones/MARCO%20JURIDICO%20INTERAMERICANO%20DEL%20DERECHO%20A%20LA%20LIBERTAD%20DE%20EXPRESION%20ESP%20FINAL%20portada.doc.pdf

preventivamente se devem ou não se expressar sobre temas de interesse público.

Esse "ostracismo digital" é especialmente grave por sua opacidade. As pessoas afetadas continuam produzindo conteúdos que praticamente ninguém vê, vivenciando um isolamento que deteriora tanto seus projetos profissionais quanto seu bem-estar emocional.

Desde a perspectiva do direito internacional dos direitos humanos, o *shadow banning* constitui uma forma particularmente problemática de restrição à liberdade de expressão. Ao se analisar essa prática sob a ótica do artigo 19 do Pacto Internacional dos Direitos Civis e Políticos (PIDCP) ou do artigo 13 da Convenção Americana, percebe-se o descumprimento dos requisitos fundamentais de legalidade, necessidade e proporcionalidade, que devem ser observados em qualquer limitação legítima a esse direito. O princípio da legalidade é comprometido porque essas restrições são implementadas com base em termos de serviço ambíguos e algoritmos opacos, que não oferecem segurança jurídica sobre quais expressões podem ser limitadas. Também não se cumprem os princípios de necessidade e proporcionalidade quando as plataformas aplicam essas medidas de forma automatizada, sem avaliar o contexto específico nem o impacto no debate público, e sem considerar alternativas menos restritivas, como avisos ou etiquetas.

A Relatoria Especial para a Liberdade de Expressão da Comissão Interamericana de Direitos Humanos (RELE) estabeleceu claramente que qualquer restrição deve ser específica e aplicada por meio de decisões fundamentadas que permitam responsabilização posterior. O *shadow banning*, por definição, viola esses princípios ao impor restrições sem notificação, explicação ou possibilidade efetiva de contestação. Além disso, baseia-se em termos de serviço confusos e imprecisos, o que levanta a questão de saber se o *shadow banning*, além de violar o princípio da legalidade, configura uma forma de censura prévia e uma restrição indireta ao direito à liberdade de expressão. A RELE já afirmou que esse direito "não pode ser objeto de medidas de controle preventivo ou prévio, mas sim da imposição de responsabilidades posteriores a quem abusar de seu exercício".[29]

A responsabilidade das grandes plataformas tecnológicas frente ao *shadow banning* é inegável. Os Princípios Orientadores sobre Empresas e Direitos Humanos da ONU (UNGP, na sigla em inglês) estabelecem que as empresas têm a responsabilidade de respeitar os direitos humanos, independentemente da capacidade ou da vontade dos Estados de cumprir com suas obrigações.[30] Para gigantes tecnológicos como Meta, TikTok ou X, isso implica três obrigações concretas: 1) identificar e prevenir impactos negativos sobre os direitos humanos, 2) implementar processos adequados conforme sua escala

29 CIDH. (2009). *Marco jurídico interamericano del derecho a la libertad de expresión*, párr. 91, p. 31.
30 Escritório do Alto Comissariado das Nações Unidas para os Direitos Humanos. (2011). *Princípios Orientadores sobre Empresas e Direitos Humanos: Implementação do quadro das Nações Unidas para "proteger, respeitar e remediar"*. Princípio 11, p. 15.
https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_sp.pdf

e, 3) fornecer mecanismos eficazes de reparação para as vítimas.

A RELE foi enfática ao destacar que:

> O exercício do poder normativo de moderação por parte das plataformas de internet — especialmente as grandes plataformas — deve estar alinhado com os princípios dos direitos humanos, o fomento ao debate público e o fortalecimento da democracia nas Américas. Elas não devem apenas aderir às normas do sistema interamericano, mas também ajustar seu poder a padrões de transparência e prestação de contas, baseados na igualdade e na não discriminação. Isso é crucial para criar um ambiente online que respeite os direitos humanos, seja livre, aberto e inclusivo, e que promova a autonomia e os direitos das pessoas usuárias.[31]

A lacuna entre o discurso corporativo e a realidade é preocupante. Enquanto Adam Mosseri, do Instagram, afirmava em 2021 que "se algo torna seu conteúdo menos visível, você deve saber disso e poder recorrer", a experiência cotidiana de criadores, ativistas e jornalistas latino-americanos revela a ausência sistemática de notificações e de mecanismos eficazes de revisão quando seu conteúdo é restringido.

Essa contradição evidencia a urgência de marcos regulatórios que reconheçam as plataformas digitais como espaços de debate público, onde se desenvolvem as discussões democráticas contemporâneas. Em uma região onde a concentração midiática historicamente limitou a pluralidade de vozes, as redes sociais inicialmente representaram uma promessa de democratização. No entanto, essa promessa está hoje comprometida por práticas opacas de moderação que reproduzem — e até amplificam — exclusões preexistentes, ainda que o façam sob o véu de decisões algorítmicas aparentemente neutras e tecnocráticas.

O que está em jogo, além das métricas e do alcance, é o direito das sociedades democráticas a um ecossistema de informação diverso, plural e acessível, no qual as vozes tradicionalmente marginalizadas possam participar em condições de igualdade na construção do debate público latino-americano.

A contradição torna-se ainda mais evidente ao contrastar as promessas corporativas com os termos e condições das plataformas. Enquanto o X menciona explicitamente em suas políticas que "limitará a visibilidade" de certos conteúdos e afirma que os usuários afetados serão notificados e poderão solicitar uma revisão, o que foi apurado nesta pesquisa indica que essas garantias raramente se concretizam. A Meta, por sua vez, admite abertamente que reduz a distribuição de "conteúdo problemático", sem se comprometer a informar as pessoas afetadas ou oferecer um mecanismo claro de apelação.

O que acontece quando as garantias prometidas nos termos e condições se tornam letra morta diante da experiência real dos usuários? Como as pessoas, em um contexto como o da América

31 CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet*, p. 61, párr. 264.

Latina, podem defender seus direitos quando nem sequer sabem que estão sendo restringidas? Quem fiscaliza se as decisões algorítmicas não estão reproduzindo vieses contra vozes historicamente marginalizadas? Essas perguntas seguem sem resposta, enquanto as pessoas navegam às cegas em um ecossistema digital no qual as regras escritas raramente coincidem com as práticas aplicadas — e onde a promessa de maior liberdade de expressão cede espaço a uma nova forma de exclusão digital tão eficaz quanto invisível.

Um debate público robusto e pluralista — essencial em qualquer sociedade democrática — exige que tanto os Estados quanto as grandes plataformas tecnológicas enfrentem essa forma invisível de silenciamento. As pessoas devem ter o direito de saber quando e por que seu conteúdo está sendo restringido, contar com mecanismos eficazes para contestar essas decisões e ter garantias de que quaisquer limitações à sua liberdade de expressão estejam de acordo com os padrões internacionais de direitos humanos. Sem essas salvaguardas, o *shadow banning* continuará a enfraquecer silenciosamente a vitalidade do debate democrático na região, criando a ilusão de participação enquanto formas sofisticadas de controle sobre o discurso público são exercidas.

05.

ALGUNS EXEMPLOS DE CASOS DE SHADOW BANNING

Com o objetivo de compreender em profundidade como o *shadow banning* opera nas plataformas digitais, esta pesquisa inclui a análise de uma série de casos representativos da América Latina e do sul global. Trata-se de situações em que jornalistas, ativistas e projetos de comunicação e incidência social denunciaram uma redução significativa no alcance, na visibilidade ou na habilidade de seus conteúdos e contas de serem descobertos, sem receber explicações claras ou notificações por parte das plataformas.

Um dos principais eixos de análise será o tratamento que Facebook, Instagram e X deram às contas e publicações que divulgam informações em apoio à causa palestina, no contexto da ofensiva militar de Israel contra a população daquele território — amplamente denunciada como um genocídio por especialistas, governos, agências da ONU e pelo Tribunal Penal Internacional.[32] Diversas denúncias públicas e estudos documentaram a aplicação de mecanismos de moderação opacos que limitam o alcance desses conteúdos, o que representa uma ameaça ao pluralismo informativo e ao direito à livre expressão nos ambientes digitais. No contexto desta pesquisa, entrevistamos Mona Shtaya, Gerente de Campanhas e Parcerias (MENA) e Líder de Engajamento Corporativo na organização Digital Action, que utiliza sua conta no Instagram para divulgar informações sobre a situação na Palestina e denunciou publicamente as restrições enfrentadas por esses conteúdos nas plataformas digitais.

Outro caso documentado é o do veículo mexicano *Chiapas Sin Censura*, que sofreu uma drástica redução de alcance no Facebook após receber uma infração por "fraude" associada a uma publicação feita faz anos. A sanção —sem aviso prévio nem possibilidade efetiva de apelação— afetou tanto a visibilidade quanto a monetização do meio, forçando sua equipe a limitar a publicação de certos conteúdos por medo de novas penalizações.

Juntamente com esse caso, também serão destacadas outras experiências significativas, como a da ativista canábica chilena conhecida pelos projetos Santiago Verde e Muy Paola, que relatou repetidas restrições de visibilidade no Instagram. No Peru, a conta Emma y Yo — um espaço de educação sexual liderado por Alesia — também denunciou uma queda abrupta no alcance de suas publicações, especialmente aquelas relacionadas a direitos sexuais e reprodutivos.

Além disso, serão analisados casos ocorridos na Argentina, como o da foto jornalista e doutora em Ciências Sociais Cora Gamarnik, que evidenciou reduções drásticas no alcance de suas publicações no Facebook sem que a plataforma oferecesse mecanismos de apelação; e o do jornalista Sebastián Lacunza, que denunciou no X uma queda inexplicável no alcance de suas publicações após comentar sobre questões políticas e de mídia.

Esses casos, contextualizados em diferentes cenários temáticos e geográficos,

32 International Criminal Court. (n.d.). *Palestine*. https://www.icc-cpi.int/palestine; Nações Unidas. (26 março de 2024). *Relatora acusa a Israel de estar cometiendo un genocidio en Gaza*. https://news.un.org/es/story/2024/03/1528636

ajudam a tornar visível como o *shadow banning* afeta o direito à liberdade de expressão e a circulação de vozes diversas — especialmente quando se trata de discursos críticos ou provenientes de setores historicamente marginalizados.

Shadow banning na Palestina: censura algorítmica contra um povo

Desde o início do ataque de Israel contra Gaza, diversos usuários vêm denunciando o Instagram, Facebook, TikTok e X por limitarem a visibilidade de suas publicações em apoio à Palestina, ainda que essas postagens não sejam removidas.[33] Em diferentes redes sociais, é possível encontrar vídeos de usuários que compartilham conteúdo pró-Palestina e relatam não ter permissão para fazer transmissões ao vivo, identificar quedas drásticas nas interações ou no número de visualizações dos seus vídeos, ou receber mensagens de outros usuários avisando que não conseguem comentar em suas publicações.

"Autores, ativistas, jornalistas, cineastas e usuários afirmam que as plataformas estão ocultando publicações com *hashtags* como #FreePalestine e #IStandWithPalestine, bem como mensagens que expressam apoio aos civis palestinos assassinados pelas forças israelenses".[34]

"A empresa diz que não há viés, que todo mundo foi afetado, mas isso não é verdade. Não vimos nenhum usuário israelense reclamar de *shadow ban*, nem mesmo em relação às tendências ou ao bloqueio de comentários", afirma Nadim Nashif, fundador da 7amleh, também conhecida como Centro Árabe para o Avanço das Redes Sociais.[35] Para driblar o *shadow banning*, alguns usuários tentam enganar os algoritmos substituindo certos termos considerados pró-Palestina por outros em suas hashtags ou publicações. No caso da Meta, a censura exercida sobre conteúdos favoráveis à Palestina chegou a ser motivo de protesto por parte de quase duzentos funcionários da empresa, por meio de uma carta aberta endereçada a Mark Zuckerberg em dezembro de 2023.[36]

Em novembro de 2024, o jornalista palestino Younis Tirawi, conhecido por denunciar crimes de guerra cometidos por Israel, começou a receber mensagens de diversos usuários do X informando sobre "falhas" ao tentar segui-lo nessa rede social.[37] O meio de comunicação *Decensored News* registrou a falha em vídeo e relatou que Tirawi "de repente perdeu a maioria de seus seguidores".

33 Observacom. (2 de outubro de 2024). *Denuncian shadowban de grandes plataformas a contenidos sobre Palestina.* https://www.observacom.org/denuncian-shadowban-de-grandes-plataformas-a-contenidos-sobre-palestina/
34 Idem.
35 France 24 Español. (2 de noviembre de 2023). *Usuarios propalestinos denuncian 'shadowban' en plataformas de redes sociales* [Video]. YouTube. https://www.youtube.com/watch?v=xm9llqJjg1A&ab_channel=FRANCE24Espa%C3%B1ol
36 *Dear Mark Zuckerberg and Leadership.* https://metastopcensoringpalestine.com/
37 Observacom. (28 de noviembre de 2024). *¿Fallas o censura oculta? polémica por restricciones a periodista palestino en X.* https://www.observacom.org/fallas-o-censura-oculta-polemica-por-restricciones-a-periodista-palestino-en-x/

Muitos deles afirmaram que deixaram de segui-lo "involuntariamente e que o X não permitia segui-lo novamente": ao clicar no botão "seguir", voltavam automaticamente a constar como não seguidores.

Embora o X tenha alegado se tratar de um "problema técnico", é altamente significativo que a mesma situação tenha ocorrido um mês antes com o perfil @Palestinahoy01 na mesma plataforma. Durante sete dias, a rede social de Elon Musk não permitia que pessoas usuárias seguissem a conta, e o número de seguidores oscilava constantemente.

A situação se repetiu nas palavras de Mona Shtaya —especialista palestina em direitos digitais e moderação de conteúdos, atual Gerente de Campanhas e Parcerias para o Oriente Médio e Norte da África e Líder de Engajamento Corporativo na organização Digital Action— que compartilhou sua experiência como usuária do Instagram afetada pelos mecanismos de *shadow banning*: "Minha conta tem mais de vinte mil seguidores e é focada principalmente na responsabilização das big techs e nos direitos digitais. Há dois anos, tenho me empenhado fortemente na divulgação da situação da Palestina."[38]

Segundo Mona, o primeiro registro documentado de *shadow banning* relacionado à Palestina ocorreu em maio de 2021, durante os acontecimentos nos bairros de Sheikh Jarrah e Silwan, em Jerusalém, quando o exército israelense tentou forçar o deslocamento de famílias palestinas. Na mesma época, houve também uma ofensiva de doze dias contra a Faixa de Gaza. Foi nesse contexto que começaram a surgir denúncias de restrições à visibilidade de conteúdos pró-Palestina nas plataformas digitais.

Com relação à sua própria conta no Instagram, Mona relatou que o *shadow banning* começou a ser ativado em novembro de 2023, aproximadamente um mês após o início da atual ofensiva militar de Israel sobre Gaza. Até então, ela não tinha enfrentado restrições semelhantes.

Mona percebeu algo estranho quando começou a comparar os números em suas estatísticas. Por exemplo, ao publicar uma *selfie*, suas *stories* podiam chegar a mais de 2.600 visualizações; já uma *story* criticando a Meta por silenciar os palestinos não chegava a 200 visualizações — uma queda de mais de 90% na visibilidade. Mona usou diferentes estratégias para fazer testes de alcance e tentar entender o que estava acontecendo. Ela relata:

> Fiz colaborações com contas grandes, com mais de dez milhões de seguidores. Naquela semana, minhas publicações alcançaram mais de um milhão de contas, mas meus *stories* mal chegaram a duzentas visualizações. Isso mostra que há um problema. Não faz sentido ter um alcance de um milhão e tão pouca visibilidade nos *stories*.

Em alguns casos, pessoas que tentaram buscar o perfil de Instagram de Mona não conseguiram encontrá-lo, e também receberam uma mensagem de advertência ao tentar enviar uma mensagem privada: "Tem certeza de

38 Entrevista realizada por Observacom y Digital Action (2025) a Mona Shtaya, 3 de abril de 2025.

que deseja enviar uma mensagem para esta pessoa?", perguntava a plataforma no momento de clicar em "enviar".

Mona Shtaya explicou que a aplicação do *shadow ban* em sua conta teve um momento claro de início:

> Eu não estava sendo *shadow banneada* no começo do genocídio. Fiz um vídeo que se tornou viral e teve cerca de duzentas mil visualizações. Basicamente, criticava a cumplicidade da Meta no genocídio. Esse vídeo viralizou e tudo o que veio depois foi uma loucura. Tudo o que publiquei depois foi fortemente *shadow banneado*.

Antes desse momento, Mona já havia trabalhado com defensoras de direitos humanos que enfrentavam restrições semelhantes, mas nunca havia passado por essa situação em sua própria conta. Ela relata:

> Foi a primeira vez que aconteceu comigo. E depois disso, senti que qualquer conteúdo que eu compartilhasse enfrentava o mesmo problema. Porque durante o genocídio, basicamente, eu não compartilhei nada além de conteúdo relacionado ao genocídio e aos direitos digitais. E todo esse conteúdo claramente teve dificuldades para alcançar as pessoas.

Em seu testemunho, Mona Shtaya compartilhou observações próprias e também de outras pessoas usuárias que documentaram casos de *shadow banning* em plataformas da Meta, vinculados a determinadas palavras-chave e símbolos associados à causa palestina. Embora no seu caso ela não costumasse usar hashtags nos stories, explicou que muitas pessoas que o faziam experimentaram uma redução notável no alcance de suas publicações quando incluíam termos como *Palestina, Gaza, genocídio* ou *apartheid*.

Uma das evidências mais significativas conhecidas sobre esse tipo de prática foi publicada pelo *The Intercept* em outubro de 2024. Segundo essa investigação, a Meta teria aplicado restrições de visibilidade a publicações que incluíam o triângulo vermelho — símbolo que muitos usuários passaram a usar para representar apoio à Palestina — sem que a plataforma informasse explicitamente sobre essa política.[39] Mona apontou que, embora já existissem suspeitas sobre esse comportamento, essa publicação serviu como prova concreta da aplicação de medidas de redução de visibilidade.

Outro padrão detectado teve a ver com comentários no Instagram que incluíam a bandeira da Palestina, corações com as cores da bandeira ou frases como "Free Palestine". Esses comentários apareciam ocultos ou marcados como "comentário oculto", sem que houvesse um aviso claro sobre o motivo. Após uma investigação solicitada pela própria Mona, o *The Intercept* replicou o problema em sua conta e, ao questionar a Meta, a empresa respondeu que

39 The Intercept. (2 de outubro de 2024). Facebook and Instagram Restrict the Use of the Red Triangle Emoji Over Hamas Association. *The Intercept*. https://theintercept.com/2024/10/02/meta-facebook-instagram-red-triangle-emoji/

apenas ocultava comentários quando detectava "discurso hostil". Essa explicação foi interpretada por ativistas como um sinal de que, ao menos na prática, a Meta estava classificando termos e símbolos relacionados à Palestina como conteúdo hostil — embora sem admitir isso publicamente nem justificar de forma clara.

A plataforma com o maior percentual de usuários que afirmam ter sofrido *shadow banning*, segundo uma pesquisa realizada pelo Center for Democracy and Technology em 2022, é o Facebook (8,1%), seguido do atual X (4,1%), Instagram (3,8%) e TikTok (3,2%). Esse tipo de censura opaca tende a afetar com mais frequência e intensidade certos movimentos sociais. Além do que está acontecendo com usuários que publicam conteúdos em apoio à Palestina, essa situação também se repete com a comunidade negra, o movimento Black Lives Matter e a comunidade LGBTQIA+.

Por que as plataformas deveriam prestar atenção às denúncias de censura que exercem? De acordo com os Princípios Orientadores sobre Empresas e Direitos Humanos das Nações Unidas (PRNU),40 as empresas têm a responsabilidade de evitar violar os direitos humanos, além de identificar e lidar com os impactos que suas operações causam nesses direitos, e oferecer mecanismos eficazes de reparação às pessoas cujos direitos foram violados.

Para as empresas de redes sociais, isso inclui alinhar suas políticas de moderação de conteúdo com os padrões internacionais de direitos humanos, garantindo que as decisões de remoção de conteúdo sejam transparentes, não excessivamente amplas ou tendenciosas, e que suas políticas sejam aplicadas de maneira consistente. Embora a Meta permita uma quantidade significativa de conteúdo pró-Palestina, isso não justifica as restrições indevidas sobre conteúdos pacíficos de apoio à Palestina, que vão contra os direitos universais à liberdade de expressão e ao acesso à informação.

Sebastián Lacunza: penalização invisível no X

Entre o final de 2022 e setembro de 2023, o jornalista argentino Sebastián Lacunza começou a notar uma queda abrupta nas interações da sua conta na rede social X. As estatísticas que ele costumava consultar ocasionalmente mostravam números anormalmente baixos: perda contínua de seguidores e pouquíssimas reações, mesmo quando era citado por usuários com muitos seguidores.

"Havia uma queda muito abrupta em todas as estatísticas e eu tinha uma perda permanente de seguidores. Às vezes, até via que pessoas com muitos

40 Escritório do Alto Comissariado das Nações Unidas para os Direitos Humanos. (2011). *Princípios Orientadores sobre Empresas e Direitos Humanos: Implementação do quadro das Nações Unidas para "proteger, respeitar e remediar"*.
https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_sp.pdf

seguidores me citavam, mas mesmo assim o alcance era baixo, o que era estranho",[41] relatou durante a entrevista.

Lacunza descobriu depois que esse fenômeno poderia estar relacionado a uma prática conhecida como *shadow banning*, embora, naquele momento, ainda não conhecesse o termo. Foi só por volta de meados de 2023, ao ouvir o termo por meio de colegas ou ativistas, que entendeu que o que estava vivendo talvez não fosse apenas uma simples queda no engajamento, mas sim uma penalização oculta imposta pela própria plataforma.

Desorientado, Lacunza chegou a considerar pagar a assinatura premium, embora não tivesse certeza de que isso resolveria o problema. No fim, decidiu publicar um tuíte denunciando a situação.[42] "Publiquei o tuíte e em menos de 48 horas tiraram a penalização. Ou até menos de 48 horas." Menos de dois dias após a publicação da denúncia sobre o *shadow banning*, sua visibilidade começou a se normalizar. "Foi imediato", afirmou. Não houve qualquer notificação oficial por parte do X, nem uma resposta automática explicando o que havia acontecido. Simplesmente, as interações começaram a aumentar e as estatísticas, que estavam em queda há meses, passaram a se recuperar.

"No começo, a recuperação foi bastante acelerada: ganhava entre quinhentos e mil seguidores por mês. Depois, estabilizou, mas algo claramente mudou", destacou. Em retrospecto, Lacunza reconhece que a penalização invisível que sofreu só pôde ser identificada por meio de observação pessoal e comparações de alcance. "Durante os meses em que estive *shadow banneado*, meu padrão era que, se um post tivesse quinze curtidas, já era muito. E de repente, depois do tuíte, cheguei a ter uma publicação com vinte e um mil curtidas. Nunca na minha vida tinha tido isso", contou. Apesar da melhora, ele ainda não sabe se as limitações impostas à sua conta foram completamente removidas ou se ainda existe algum tipo de restrição parcial.

Outro indício claro do *shadow banning* foi a perda da habilidade de sua conta ser descoberta. Sebastián relatou que, durante o período em que suspeitava estar sendo penalizado por algoritmos, era notoriamente difícil encontrar seu perfil, mesmo digitando seu nome completo na busca da plataforma. "Eu digitava meu nome na busca e minha conta não aparecia. Em vez disso, apareciam três ou quatro contas falsas que alguém criou usando meu nome e minha foto de perfil", explicou. Essa experiência coincide com um dos sinais mais mencionados por outros usuários em fóruns e tutoriais para identificar um *shadow banning*: a exclusão dos resultados de busca, mesmo quando se procura exatamente pelo nome do usuário.

Essa falta de visibilidade não se devia a um erro do sistema nem a uma falha geral no mecanismo de busca, mas afetava exclusivamente sua conta — enquanto os perfis falsos com seu nome eram facilmente encontrados. "Esse é um dado importante, porque não é que

41 Entrevista realizada por Observacom e Digital Action (2025) a Sebastián Lacunza, 18 de março de 2025.
42 Lacunza, S. (15 de setembro de 2023). Me percato que tengo shadowban de Twitter hace meses. [Posteo]. X. https://x.com/sebalacunza/status/1702735419777880281

nenhum resultado aparecia: apareciam essas contas falsas, mas a minha, não", ressaltou Lacunza.

Seu caso ilustra com clareza como o *shadow banning* pode operar de forma silenciosa, sem que os usuários sequer saibam que estão sendo afetados — e sem acesso a mecanismos claros de apelação ou revisão. Ao mesmo tempo, evidencia como essa penalização pode ser suspensa sem qualquer explicação, reforçando o caráter opaco dessas práticas e a ausência de garantias para quem tem sua visibilidade comprometida em ambientes digitais que hoje funcionam como espaços públicos de deliberação e expressão.

Shadow banning à mídia independente Chiapas Sin Censura

O meio de comunicação mexicano Chiapas Sin Censura, fundado em 2012, tem sido um caso relevante de *shadow banning*, especialmente no Facebook. Seu fundador e diretor geral, José David Morales Gómez, denunciou graves restrições de visibilidade nas redes sociais, em particular no Facebook. Segundo relato de David, o grupo sofreu uma queda drástica no alcance após receber uma sanção por uma suposta infração relacionada a uma publicação feita quatro anos antes. A plataforma classificou o conteúdo como "fraude", sem oferecer maiores explicações nem a possibilidade de apelar de forma efetiva. A publicação em questão divulgava o pedido de um jovem com câncer que solicitava apoio ao boxeador Canelo Álvarez, conhecido por suas ações solidárias.

"Essa publicação, que era real, também estava circulando em outras páginas e não tinha acontecido nada. De repente, recebemos a infração por fraude, e fomos diretamente do verde para o vermelho", relatou David. Até aquele momento, Chiapas Sin Censura acumulava mais de cem milhões de visualizações mensais; após a penalização, o alcance caiu para vinte e oito milhões. "Nosso crescimento orgânico parou, os seguidores começaram a cair, perdemos a monetização e não obtivemos nenhuma resposta às nossas apelações", explicou.

Após receber a infração, os jornalistas tentaram recorrer da decisão por diversas vias, sem obter respostas. David explicou que utilizou a opção de apelação disponibilizada pela própria plataforma no momento da notificação da sanção: "Dizia que iriam nos responder em quatro dias úteis, mas nunca responderam." Ele também enviou e-mails, abriu relatórios manuais a partir do perfil de administrador e se inscreveu no serviço Meta Verify — uma ferramenta paga que promete atendimento personalizado —, com a esperança de obter uma resposta mais ágil. A assinatura custava cerca de sete mil pesos mexicanos por mês.

Durante esse processo, conseguiu falar uma vez com uma pessoa da Meta, que informou que o caso já havia sido encaminhado e que ele receberia uma resposta em 24 horas. Essa foi a única interação direta que teve. A partir daí, não houve mais contato. "Tentava fazer

outro relatório e aparecia que já havia uma conversa iniciada. Não foi possível fazer mais nada", relatou.

A falta de vias efetivas para recorrer foi um dos aspectos mais frustrantes de todo o processo. A sanção permaneceu ativa por semanas, afetando diretamente o alcance, a monetização e o vínculo com o público. "Se não fosse pelo apoio de uma organização que interveio para nos ajudar, acredito que a sanção teria durado um ano", afirmou David, em referência ao acompanhamento recebido.

Esse caso evidencia as limitações dos canais internos de apelação da Meta, o que afeta diretamente o direito dos usuários de se defender. A ausência de uma instância clara de revisão — acessível, transparente e com prazos razoáveis de resposta — representa uma vulnerabilidade crítica para as mídias que dependem quase exclusivamente das plataformas digitais para sua distribuição e sobrevivência. Isso também impacta negativamente o direito da população de receber informação, especialmente aquela que se informa por meio desses canais.

O impacto não foi apenas técnico ou econômico: também prejudicou o trabalho jornalístico. Segundo David, passaram a evitar publicar temas sensíveis — como casos de violência, crianças em situação de vulnerabilidade ou denúncias contra o crime organizado — por medo de novas penalizações. "Às vezes dizemos: essa matéria não vale uma infração. Mesmo que o fato seja real, preferimos não publicar, porque pode nos prejudicar ainda mais."

O bloqueio teve ainda um efeito emocional e editorial."Passei dias sem vontade de publicar. Pra quê fazer isso, se o trabalho não chega às pessoas?", desabafou. Assim como ocorre com muitos outros meios digitais na região, o principal canal de Chiapas Sin Censura é o Facebook. "Se amanhã decidirem apagar nossa página, perdemos doze anos de trabalho e dez famílias ficam sem renda."

Apesar de a sanção ter sido retirada no fim, o episódio escancara a falta de mecanismos acessíveis de defesa diante de decisões automatizadas e unilaterais. Também revela como a ameaça de invisibilização impacta diretamente a cobertura de temas de interesse público e a sustentabilidade de mídias locais e críticas.

@muypaola: quando falar sobre cannabis te torna invisível nas redes

Diante do celular, durante a vídeo chamada da entrevista, Paola Díaz mostra as estatísticas da sua conta no Instagram: "Sua conta não pode ser mostrada para pessoas que não te seguem",43 lê em voz alta a notificação, sem maiores explicações. Com frustração, aponta a diferença brutal no alcance entre publicações aparentemente semelhantes: um conteúdo ativista recebe 5.300

43 Entrevista realizada por Observacom e Digital Action (2025) a Paola Díaz, 12 de março de 2025.

visualizações, enquanto outro, com temática festiva, chega a 70 mil.

A ativista canábica chilena, criadora dos perfis @stgoverde44 e @muypaola, denuncia a redução sistemática de seu alcance, sem qualquer notificação explícita ou justificativa transparente. A conta não é bloqueada diretamente, mas se torna praticamente invisível. "O *shadow ban* começa a mostrar cada vez menos, e só para os seus seguidores", explica Paola, enquanto mostra como suas postagens — que antes alcançavam centenas de milhares de visualizações — agora chegam apenas ao seu círculo mais próximo. Até sua capacidade de ser encontrada nas buscas foi severamente limitada: "Quando me procuram, têm que digitar o nome completo e apertar 'enter', porque nem aparece nas sugestões", acrescenta, descrevendo um sistema que parece ter sido projetado para fazer com que certos perfis simplesmente desapareçam do radar público.

O caso de Paola ilustra como a intervenção algorítmica de plataformas como o Instagram pode operar com uma dupla invisibilidade: oculta o conteúdo do restante dos usuários e oculta esse processo do próprio criador. Ao contrário de uma suspensão direta, o *shadow banning* não vem acompanhado de notificações formais, nem de explicações sobre quais termos ou temas estão sendo penalizados. Paola, após anos de documentação, conseguiu identificar padrões específicos. Termos como *maconha, 420*, ou até informações sobre a cultura canábica e os direitos relacionados, acionam filtros invisíveis que reduzem drasticamente o alcance. "A Meta identifica palavras que não podem ser ditas porque senão você leva *shadow ban*", revela, expondo um sistema de moderação temática altamente específico.

Essa moderação oculta teve consequências devastadoras que vão muito além da visibilidade online. Em setembro de 2024, após anos de *shadow banning* intermitente, o Instagram finalmente desativou uma de suas principais contas depois de exigir uma verificação biométrica. O mais impressionante: "Continuaram cobrando a assinatura [do Meta Verified]", conta, indignada, ao descrever como a Meta continuou faturando por um serviço de verificação em contas que ela já não podia mais usar. Mais uma vez, a falta de transparência se combina com práticas comerciais questionáveis, sem que haja mecanismos eficazes de recurso ou reparação.

O impacto emocional e profissional, assim como no caso de *Emma y Yo*, foi profundo para Paola. "Agora, quando entro nas redes sociais, começo a ter crises de ansiedade", confessa. O medo constante do ostracismo digital e de ser silenciada gerou uma autocensura preventiva que frustra sua capacidade de educar sobre temas cruciais: "Não posso falar sobre redução de danos. Muitas vezes eu gostaria de falar sobre problemas do narcotráfico, [mas] não consigo alertar minha comunidade sobre riscos reais." O temor de ser expulsa de uma plataforma tem um efeito inibidor sobre seu direito de participar do debate sobre a cannabis no Chile. A paradoxo

44 No momento desta publicação, a conta @stgoverde permanece suspensa pelo Instagram, após múltiplos ciclos de restrição e recuperação. A plataforma não apenas suspendeu a conta, mas também havia removido sistematicamente conteúdos de interesse público, incluindo reportagens jornalísticas sobre a cultura canábica e material educativo produzido pela ativista.

é evidente: enquanto as plataformas justificam suas políticas de moderação como forma de proteger contra danos, estão silenciando justamente as vozes que poderiam prevenir riscos concretos relacionados ao consumo e à criminalização da cannabis.

A resposta de Paola frente a essa situação tem sido multidimensional, combinando estratégias jurídicas e ações coletivas de visibilização. Em parceria com um escritório jurídico, ela apresentou um recurso de proteção no Chile, que foi rejeitado e agora está em processo de apelação. Ao mesmo tempo, denunciou essas práticas ao Serviço Nacional do Consumidor (Sernac), argumentando que há uma violação dos direitos do consumidor, já que a Meta fatura no Chile sem respeitar a jurisdição local. "Eles estão faturando no Chile, portanto, devem atuar sob a jurisdição chilena", afirma, levantando um desafio regulatório que ultrapassa fronteiras nacionais.

O caso de Paola é exemplo de uma problemática global que afeta diversas comunidades: desde ativistas canábicos até defensoras de direitos humanos na Palestina, passando por criadoras de conteúdo de moda inclusiva com corpos não hegemônicos. Paola tem buscado formar redes transnacionais com pessoas afetadas com o objetivo de evidenciar a natureza sistemática dessa censura seletiva. "Tenho um grupo com muitas pessoas de diferentes partes do mundo — Argentina, Tailândia, Espanha, México, Uruguai — todos canábicos, e todos tiveram suas contas derrubadas", explica, descrevendo como criadores com centenas de milhares de seguidores vêm enfrentando o mesmo padrão de invisibilização progressiva seguida de exclusão completa das redes.

Cora Gamarnik: suspensão e queda drástica no alcance no Facebook

"O que senti foi uma redução drástica no alcance e nas interações da minha conta do Facebook, especialmente a partir da publicação que compartilhei sobre o caso de Lago Escondido, onde postei um print da conversa entre os juízes", explica a doutora em Ciências Sociais Cora Gamarnik. Em 12 de dezembro de 2022, Gamarnik fez uma publicação no Facebook referindo-se à viagem de juízes, funcionários públicos e membros do Grupo Clarín, na Argentina, e às mensagens trocadas entre eles, que haviam se tornado públicas naquela mesma semana.

O Facebook removeu sua publicação e, em seguida, bloqueou sua conta temporariamente. "Quando publiquei aquilo, suspenderam minha conta alegando que eu estava disseminando discurso de ódio. Então, fiz uma reclamação e expliquei que meu post era justamente contra as mensagens de ódio presentes na captura de tela que compartilhei", relata ela. A suspensão durou alguns dias.

Sua publicação consistia em uma série de prints de tela da conversa, onde apareciam frases como "vamos limpar um mapuche", acompanhadas de um texto escrito por ela denunciando o racismo

presente naquele caso. “Claramente o Facebook não leu o texto. Depois disso, meus posts passaram a ter um alcance muito limitado, bem diferente do que acontecia antes”, explica a pesquisadora do Conicet.

Ao notar a queda no alcance de suas postagens — com muito menos interações do que o habitual —, Gamarnik passou a usar menos o Facebook e a focar sua atividade em outras redes, embora nunca tenha abandonado completamente a plataforma. Ela explica:

> Até aquele momento, minha conta no Facebook tinha muita difusão, e eu sabia com certeza que, se publicasse determinadas coisas, elas se tornariam virais ou seriam compartilhadas imediatamente. O que comecei a perceber é que essas mesmas postagens passaram a não ter resultado algum.

Além disso, muitas pessoas que a seguiam e liam com frequência começaram a dizer que não viam mais suas postagens, e até pensaram que ela tinha parado de usar o Facebook.

Atualmente, ao observar a evolução de suas métricas, é possível ver como as interações nas publicações de Cora Gamarnik caíram significativamente a partir de 12 de dezembro de 2022 — passando de posts com mais de dois mil curtidas, dezenas de comentários e mais de 300 compartilhamentos (no início de dezembro daquele ano), para publicações com apenas 13 curtidas, 2 compartilhamentos e nenhum comentário.

O caso de Cora Gamarnik se soma ao de outras figuras públicas, como o do jornalista Sebastián Lacunza, que, em setembro de 2023, publicou em sua conta no X (antigo Twitter) que havia percebido estar sofrendo um *shadow banning*, já que: “As ‘impressões’ caíram para um quinto do normal, e as interações despencaram abruptamente.” Além disso, ele afirmou que deixou de aparecer nos resultados de busca para quem não o seguia, o que teve um impacto enorme no alcance e na visibilidade de sua conta.

Emma y Yo: quando reduzem o alcance da educação integral em sexualidade no Instagram

“Estive a ponto de jogar tudo pro alto e dizer: já era, foda-se... que o material fique aí, que lindo, acabou, não quero mais”.[45] A frustração nas palavras de Alesia Lund, criadora do projeto de educação sexual @emmayyoperu[46] no Instagram, revela a realidade invisível enfrentada por educadoras digitais na América Latina. Sem alertas, sem explicações, ela viu sua conta — que chegou a ter quase 68 mil seguidores — desaparecer sistematicamente diante de seus olhos, perdendo mais de 10 mil seguidores em dois anos, enquanto suas

45 Entrevista realizada por Observacom e Digital Action (2025) a Alesia Lund, 14 de março de 2025.
46 Emma y Yo Perú. (14 de abril de 2025). Instagram. *https://www.instagram.com/emmayyoperu/*

publicações passaram de 15 a 20 mil visualizações para apenas 300 ou 500. O que Alesia descreve se encaixa no fenômeno que esta investigação vem documentando: uma prática de moderação de conteúdo que age como um fantasma no algoritmo — sem deixar rastros, sem enviar notificações, apenas apagando o conteúdo do radar público.

O projeto *Emma y Yo* nasceu em 2019, como uma resposta urgente à escassez de conteúdo acessível sobre educação sexual no contexto latino-americano. Através de ilustrações, infográficos e materiais didáticos cuidadosamente elaborados, o projeto rapidamente se tornou uma referência regional para falar sobre sexualidade com crianças, adolescentes e jovens adultos. Combinando informação de qualidade com uma estética acessível e atrativa, o projeto desmistifica tabus — da anatomia ao consentimento — usando uma linguagem direta e inclusiva. Com dois livros publicados e uma comunidade majoritariamente composta por mulheres jovens entre 18 e 24 anos, a iniciativa de Alesia preencheu um vazio educativo crítico em um continente onde a educação sexual institucional ainda é insuficiente. O crescimento orgânico do projeto nas redes refletia não só a qualidade do conteúdo, mas também a necessidade social urgente por informação confiável sobre um tema historicamente silenciado.

O mais revelador no caso de Alesia é a ausência total de comunicação por parte da plataforma. Diferente de uma suspensão explícita, que inclui notificações por violação de diretrizes, Alesia nunca recebeu nenhum aviso sobre seu conteúdo educativo. Essa invisibilização progressiva aconteceu em paralelo às mudanças nas políticas de conteúdo da Meta, especialmente no período pós-pandemia, quando várias contas educativas sobre sexualidade começaram a relatar problemas semelhantes. “Comecei a ver contas gringas que eu seguia sendo banidas. Vi de longe e pensei: essa merda vai chegar aqui também.” E chegou. Pouco tempo depois, Alesia começou a perder seguidores e suas publicações passaram a ter cada vez menos visualizações. O padrão se confirma quando comparado com outras criadoras: conteúdos sobre temas considerados controversos — como educação sexual, feminismo ou direitos reprodutivos — têm muito menos alcance do que outros tipos de conteúdo.

O impacto dessa moderação invisível, opaca e silenciosa vai muito além dos números: afeta emocional e profissionalmente. “Minha produção caiu muito”, confessa Alesia, descrevendo ciclos de frustração que a levaram a pausar o trabalho por semanas. “Pra que diabos eu me esforço, se ninguém vê o que eu faço?”, perguntava a si mesma. Essa erosão gradual de alcance não impacta apenas a motivação criativa — compromete diretamente a sustentabilidade de projetos educativos independentes, forçando criadores a migrar constantemente entre plataformas (de Facebook para Instagram, depois para TikTok e YouTube) em busca de visibilidade. É uma situação que, nas palavras de Alesia, desgasta profundamente. Por isso, ela tem buscado se afastar do Instagram e focar em plataformas como o YouTube, na esperança de “finalmente monetizar seu projeto pela primeira vez”.

O caso evidencia como o *shadow banning* funciona como um mecanismo de controle sem transparência, que afeta de forma desproporcional vozes que

abordam temas considerados sensíveis. O algoritmo parece penalizar palavras como "sexualidade", "vulva" ou "educação sexual", mesmo quando usadas em contextos educativos. Por isso, Alesia passou a evitar certos *hashtags*. O mais preocupante é que, ao contrário da censura tradicional, esse sistema não oferece qualquer possibilidade de apelação: sem notificações, sem saber que regras foram supostamente violadas, os criadores não conseguem ajustar seu conteúdo para seguir diretrizes específicas. Essa forma de controle invisível representa um desafio fundamental à liberdade de expressão nos ambientes digitais, especialmente para quem aborda temas de saúde sexual a partir de perspectivas educativas e feministas na América Latina.

O depoimento de Alesia também aponta para um fenômeno mais amplo que afeta contas de educadores em toda a região e que está associado ao *shadow banning*: a comercialização da atenção em plataformas como o Instagram está exercendo uma pressão crescente para monetizar o alcance. "Estão nos empurrando para que tenhamos que pagar", afirma Alesia, explicando como até mesmo contas novas, sem conteúdo considerado polêmico, enfrentam limitações severas se não investem em promoção. "Antes o crescimento era orgânico, rápido. Agora o crescimento é lentíssimo, as visualizações não vêm", relata. Essa realidade levanta questionamentos sobre o acesso democrático à informação, especialmente quando conteúdos educativos sobre sexualidade, de interesse público, são restringidos por algoritmos comerciais que não distinguem entre conteúdo sensível e conteúdo educativo — criando, assim, um duplo padrão particularmente prejudicial para iniciativas sem fins lucrativos.

Por fim, os padrões descritos por Alesia sugerem a existência de uma forma de discriminação algorítmica baseada na temática. A forma como as plataformas priorizam certos tipos de conteúdos — como vídeos e imagens "leves" — em detrimento de ilustrações educativas ou textos informativos, revela uma lógica em que determinados formatos e temas são privilegiados em detrimento de outros. Esse viés técnico intensifica ainda mais a invisibilização, pois mostra, por um lado, que vídeos de entretenimento sobre sexualidade circulam livremente, enquanto, por outro, materiais infográficos com abordagem científica e pedagógica acabam relegados ao esquecimento.

Quando os algoritmos julgam corpos: o caso Love&Lust contra a invisibilização digital

Essa discriminação algorítmica e o *shadow banning* não afetam apenas o conteúdo educativo ligado à sexualidade. Empreendedoras independentes, que constroem comunidades em torno de produtos considerados "sensíveis", enfrentam o mesmo muro invisível de silenciamento. O caso de Paula Labra, com seu negócio de lingerie inclusiva, é mais um exemplo claro desse fenômeno.

"Gravei uma amiga procurando minha conta e simplesmente ela não existia",[47] relatou em uma entrevista. Com essas

47 Entrevista realizada por Observacom e Digital Action (2025) a Paula Labra, 19 de março de 2025.

palavras, Paula, proprietária do e-commerce de lingerie @lovelust.cl, com 240 mil seguidores, descreve o momento em que documentou aquilo que milhares de criadores suspeitam, mas raramente conseguem provar: sua conta havia sido submetida a *shadow banning*, tornando-se praticamente invisível na plataforma que sustenta boa parte de seu negócio.

O caso da *Love&Lust* expõe com clareza os mecanismos opacos de moderação que afetam desproporcionalmente criadores de conteúdo. Paula relata como o Instagram e o TikTok aplicam critérios discriminatórios baseados no tipo de corpo. "No TikTok, se era uma mulher magra, não derrubavam. Se era uma mulher voluptuosa, sim. Um pouco mais gordinha ou com mais [peito], tchau, derrubavam o post." Esse viés algorítmico revela uma forma de censura que vai além das políticas explícitas e reproduz preconceitos contra corpos não normativos, afetando especialmente conteúdos que celebram a diversidade corporal feminina ou mostram produtos relacionados à anatomia, como próteses de mamilos pós-mastectomia ou roupa íntima menstrual.

Os efeitos do *shadow banning* são devastadores e mensuráveis. "Senti isso nas vendas. Era como se não existíssemos, mesmo sendo uma marca que investia três mil dólares por mês em publicidade", relata Paula. Seu testemunho mostra como a Meta cobra por serviços publicitários enquanto simultaneamente limita o alcance das contas que pagam por visibilidade — um paradoxo que, como veremos adiante, também se estende a outros serviços da plataforma. Essa situação criou uma nova forma de ilusão algorítmica, em que criadores pagam 45 dólares mensais por uma assinatura do Meta Verified "por medo", sem qualquer garantia real de proteção contra a invisibilização. "Estou gastando mais de quinhentos dólares por ano por medo. Isso é uma estupidez", lamenta. O impacto psicoemocional também é significativo: "Vivo com medo. Antes de dormir, sempre peço que minha conta não seja derrubada." Essa ansiedade constante afeta diretamente seu bem-estar e sua capacidade criativa.

A resposta de Paula a essa censura sistemática mostra a resiliência engenhosa de criadoras latino-americanas diante de algoritmos opacos. Ela desenvolveu estratégias comunitárias, organizando um grupo de amigas influenciadoras para repostarem constantemente seus conteúdos, tentando romper o *shadow banning* através de interações em massa. Ao mesmo tempo, documenta meticulosamente cada episódio de censura, salvando capturas de tela que evidenciam o tratamento desigual entre contas pequenas e grandes marcas, como Calvin Klein ou Savage X Fenty, que conseguem mostrar conteúdo muito mais explícito sem sofrer penalizações. O caso de Paula mostra como, mesmo não se tratando de um projeto com foco direto em temas de interesse público, a moderação algorítmica está moldando um ecossistema digital onde certos corpos e assuntos são sistematicamente invisibilizados. E obriga pequenos empreendedores a desviar recursos — que poderiam ser usados para melhorar seus produtos — para lutar às cegas contra um sistema opaco, que submete sua existência a um medo constante de desaparecer.

06.

SHADOW BANNING NAS REGRAS DAS PLATAFORMAS

Um dos objetivos desta investigação foi analisar quais informações as plataformas digitais oferecem aos seus usuários sobre as medidas de moderação de conteúdo que afetam o alcance e a visibilidade de suas publicações ou contas — conhecidas como *shadow banning* —, bem como as possibilidades de recorrer ou contestar essas decisões.

Nesta seção, são examinados os termos e condições da plataforma X e da Meta (empresa-mãe do Facebook e Instagram, cujas regras são unificadas), além de declarações de seus diretores e porta-vozes, para identificar como essas plataformas definem, explicam e regulam a redução de alcance ou visibilidade, e que tipo de garantias e direito de defesa oferecem, se é que oferecem, às pessoas usuárias afetadas.

Termos e condições da plataforma X

No caso das regras da plataforma X, há uma seção específica sobre as medidas que podem ser aplicadas quando um conteúdo é considerado violador das normas da plataforma. Ali, são diferenciados três níveis de ação: medidas no nível da publicação (post), medidas no nível da conta e medidas aplicadas a mensagens diretas.

Entre as ações que a plataforma pode aplicar sobre uma publicação, está expressamente prevista a possibilidade de "limitar a visibilidade do post". Nas suas normas, a plataforma explica que toma medidas

> quando um post específico viola as Regras do X, inclusive os X que compartilha ou reproduz outros X ao publicar capturas de tela, posts com comentários ou compartilhamento de URLs de posts que violam nossas regras.

A limitação de visibilidade é descrita da seguinte forma: "quando adequado, restringiremos o alcance dos posts que violarem nossas políticas e criaremos uma experiência negativa para outros usuários ao deixar os posts menos visíveis no X". As medidas possíveis incluem:

1. Excluir o post dos resultados de pesquisa, assuntos e notificações recomendadas.
2. Remover o post das timelines Para Você e Seguindo.
3. Restringir a visibilidade do post para o perfil do autor.
4. Rebaixar o post nas respostas.
5. **Restringir** curtidas, respostas, Reposts, posts com comentários, Itens Salvos, compartilhamentos, posts fixos no perfil ou post com edição.

A partir de abril de 2023, a plataforma passou a etiquetar publicamente os posts identificados como infratores de suas normas, informando tanto o autor quanto os leitores de que a visibilidade daquele conteúdo foi limitada. Os autores podem solicitar uma revisão dessa etiqueta caso considerem que a limitação foi aplicada incorretamente. No entanto, não há uma explicação clara

sobre como esse pedido deve ser feito — ao contrário dos procedimentos detalhados que existem para recorrer de suspensões ou bloqueios de contas.48

Além disso, a X prevê uma "exceção por interesse público"49 para certos conteúdos que, embora infrinjam as regras, devem permanecer acessíveis por sua relevância pública. Essa exceção é aplicada principalmente a publicações feitas por contas de alto perfil que representam membros atuais ou potenciais de órgãos governamentais ou legislativos. Nesses casos, o post é colocado atrás de um aviso, sua visibilidade é limitada, mas ele permanece disponível na plataforma.

Os critérios para aplicar essa exceção são:

- O post infringe uma ou mais regras da plataforma;
- O post foi publicado por uma conta de alto perfil;
- A conta representa um membro atual ou potencial de um órgão legislativo ou governamental, em nível local, estadual, nacional ou supranacional. Isso inclui:
  - **a.** Titulares de cargos de liderança eleitos ou indicados;
  - **b.** Candidatos ou indicados a cargos políticos;
  - **c.** Partidos políticos registrados.

Termos e condições da Meta (Facebook e Instagram)

A Meta, empresa controladora do Facebook, Instagram, Threads e Messenger, inclui em seus documentos oficiais referências explícitas à prática de reduzir a visibilidade de certos conteúdos, mesmo quando não violam suas regras da comunidade.50 Essa medida faz parte de sua política de "curadoria" de conteúdo, estruturada desde 2016 em torno de três pilares: remover, reduzir e informar.

O pilar de "remover" é o mais facilmente reconhecível, pois envolve ações clássicas de moderação, como a remoção de conteúdos e a exclusão de contas de usuários. Por outro lado, conforme explicado na seção "Reduzir a distribuição de conteúdo problemático"51 no Centro de Transparência da empresa, o pilar de "reduzir" visa limitar a circulação do que a plataforma chama de "conteúdo problemático" — ou seja, conteúdo

48 X Central de ajuda. Recorra de uma conta bloqueada ou suspensa. https://help.x.com/pt/forms/account-access/appeals/redirect
49 X Central de ajuda. Sobre as exceções devido ao interesse público no X. https://help.x.com/pt/rules-and-policies/public-interest
50 Desde 12 de novembro de 2024, a Meta unificou suas normas comunitárias em uma única norma que se aplica aos seus quatro serviços de redes sociais e mensagens: Facebook, Instagram, Messenger e Threads. As normas comunitárias podem ser lidas na íntegra aqui: https://transparency.meta.com/pt-br/policies/community-standards/
51 Meta. (25 de abril de 2025). Reduzindo a distribuição de conteúdo problemático. https://transparency.meta.com/pt-pt/enforcement/taking-action/lowering-distribution-of-problematic-content/

que, embora não viole diretamente as regras, pode "gerar experiências negativas" ou ser considerado de "baixa qualidade". Nesses casos, a Meta declara que reduz a distribuição do conteúdo no *feed* e nas recomendações, sem removê-lo e sem notificar o usuário de que foi penalizado.

A empresa mantém uma classificação ampla e ambígua do que considera "conteúdo problemático", sujeito à redução de visibilidade. Entre os exemplos citados estão:

- Conteúdo de baixa qualidade, como isca de cliques e isca de engajamento;
- Links para sites cobertos com anúncios, com carregamento lento ou quebrados;
- Comentários de baixa qualidade que são copiados e colados repetidamente;
- Conteúdo com originalidade limitada, predominantemente reaproveitado de outras fontes;
- Vídeos de baixa qualidade, que abusam de formatos como o vídeo comum ou a transmissão ao vivo;
- Informações incorretas ou desinformação;
- Conteúdo de criadores que repetidamente violam as políticas da plataforma.

Como pode ser observado, a lista inclui desde *clickbaits* e comentários repetitivos até conteúdos sem originalidade e publicações marcadas como desinformação. É importante destacar que essa última categoria pode resultar em uma penalização algorítmica sem que haja uma avaliação humana nem um canal de apelação acessível para os usuários afetados.

A Meta justifica essa prática como um mecanismo para priorizar a experiência dos usuários. No entanto, a empresa não oferece notificações nem vias de apelação acessíveis nesses casos, o que impede que os usuários compreendam se estão sendo penalizados e por quê. Essa falta de transparência é particularmente relevante quando decisões automatizadas afetam a circulação de conteúdo legítimo ou relacionado a temas de interesse público.

Além disso, as diretrizes de recomendação que a Meta aplica ao Facebook e Instagram reforçam esse tipo de prática. Nessas diretrizes, está estabelecido que o conteúdo sugerido — como aquele que aparece no "Explorar", em "Sugestões para você" ou nos reels — é regido por critérios internos destinados a evitar a amplificação de materiais que a plataforma considera inadequados ou irrelevantes para certos públicos. Contudo, não são explicitados os parâmetros exatos que orientam essas decisões, nem são oferecidos mecanismos transparentes para que os usuários compreendam por que seu conteúdo deixou de circular normalmente.

Diferentemente do que ocorre com outros mecanismos de moderação de conteúdo — como a remoção de publicações ou a suspensão de contas —, no caso da "redução de distribuição de conteúdo problemático", a Meta não prevê em suas políticas a possibilidade de apelar ou solicitar a revisão da medida. Isso aprofunda a opacidade do processo, já

que os usuários não recebem nenhuma notificação, nem dispõem de ferramentas para reverter ou contestar decisões que afetam diretamente a visibilidade de suas publicações.

Entre as regras da comunidade da Meta também estão incluídas restrições específicas ao conteúdo relacionado à cannabis e seus derivados. Na seção sobre Produtos Regulamentados, a empresa estabelece que pode restringir publicações que "coordenem ou promovam (ou seja, que falem positivamente, incentivem o uso ou forneçam instruções de uso ou fabricação) a maconha e produtos que contenham THC ou componentes psicoativos relacionados".52 Embora a política não implique necessariamente a exclusão de contas que compartilhem esse tipo de conteúdo, sua inclusão nas regras permite que a empresa aplique mecanismos de redução de visibilidade ou limitação de alcance, pelo menos entre usuários menores de 18 anos. Esse marco regulatório pode estar relacionado ao caso da conta do Instagram @muypaola, cuja criadora denunciou uma queda abrupta nas interações e na visibilidade de suas publicações. No entanto, no caso dela, a limitação de alcance não parece ter sido aplicada apenas a menores de 18 anos, mas sim a usuários em geral.

52 Meta. Produtos e serviços restritos. https://transparency.meta.com/pt-br/policies/community-standards/regulated-goods/

07.

CONCLUSÕES

Esta investigação permitiu identificar que o *shadow banning* é um conjunto de práticas cada vez mais frequentes nas principais plataformas digitais e que tem um impacto significativo sobre a liberdade de expressão, a circulação de informações de interesse público e a participação democrática nos ambientes digitais.

Por meio da análise de casos concretos na América Latina, de entrevistas com especialistas e da revisão dos termos e condições de empresas como a Meta (Facebook e Instagram) e X (antigo Twitter), ficou evidente que essas formas opacas de moderação operam como mecanismos de silenciamento algorítmico, afetando criadores de conteúdo, ativistas, jornalistas e comunidades historicamente marginalizadas. Elas ocultam conteúdos por meio de práticas que os usuários não conseguem detectar.

Um dos principais achados é que a redução de visibilidade pode ter consequências equivalentes — ou até mais duradouras — que a remoção direta de conteúdo. Mesmo que os conteúdos não sejam excluídos, sua circulação é severamente restringida, afetando tanto o alcance quanto a possibilidade de participação no debate público. Essa penalização, que muitas vezes ocorre de forma automatizada, pode gerar efeitos desproporcionais, que só são percebidos pelas pessoas usuárias após notarem uma queda repentina nas interações, visualizações ou no alcance de suas contas.

Além disso, constatou-se que a transparência das plataformas em relação a essas práticas é mínima ou inexistente. A Meta reconhece que aplica medidas para "reduzir a distribuição de conteúdo problemático" — uma categoria ampla e ambígua —, mas não informa aos usuários quando essas medidas são ativadas, nem oferece vias claras de apelação. O X, por sua vez, menciona que pode limitar a visibilidade de certos conteúdos e promete uma possível revisão, mas não detalha os procedimentos, nem garante sua acessibilidade ou eficácia. Na prática, isso deixa os usuários afetados em uma situação de vulnerabilidade, sem informações suficientes para contestar a penalização ou ferramentas para revertê-la.

A investigação também demonstra que as decisões sobre quais conteúdos terão sua visibilidade reduzida não são neutras. Os algoritmos de moderação não apenas reproduzem preconceitos sociais existentes, como também amplificam desigualdades estruturais, ao limitar o acesso a vozes dissidentes ou fora da norma dominante.

Por fim, o estudo confirma que essas práticas violam direitos individuais e também comprometem o caráter público e democrático dos espaços digitais. Se as plataformas continuarem operando sem transparência, sem prestação de contas e sem mecanismos adequados de defesa para os usuários, o risco vai além da censura encoberta de determinadas vozes: coloca em xeque o próprio debate democrático na internet.

Diante desse cenário, é urgente avançar em marcos normativos e regulatórios que garantam transparência, devido processo legal e direito à defesa diante das medidas de redução de alcance e visibilidade aplicadas pelas grandes plataformas digitais.

OBSERVACOM

www. observacom.org