7 REFLEXÕES

para o futuro do debate sobre moderação de conteúdo em plataformas digitais

AUTOR:
Carlos Affonso Souza

MARÇO 2025

ITS

No contexto da *Audiência Pública para Debate Técnico sobre a Política de Moderação de Conteúdo das Plataformas Digitais no Brasil*, realizada pela Advocacia-Geral da União, no dia 22.01.25, o **Instituto de Tecnologia e Sociedade (ITS Rio)** vem apresentar sua contribuição, constituída de 7 (sete) reflexões para o futuro do debate sobre moderação de conteúdo.

Essas reflexões endereçam questionamentos e tendências que podem ser percebidas na literatura acadêmica sobre moderação de conteúdo, bem como aspectos práticos que vêm se tornando notórios nas dinâmicas de moderação adotadas por empresas e demais entidades que administram ambientes online nos quais terceiros publicam diversas formas de conteúdo.

O ITS Rio espera que esta manifestação possa contribuir para a formatação de políticas públicas e para o aperfeiçoamento das práticas de moderação, facilitando o entendimento sobre aspectos centrais relacionados ao tema e o seu cada vez mais atual e relevante impacto na sociedade.

01.

QUEM DEVE FAZER MODERAÇÃO DE CONTEÚDO?

Moderação de conteúdo é feita por empresas (grandes e pequenas), entidades da sociedade civil e por autoridades públicas na gestão de seus ambientes online.

Moderação de conteúdo não é uma atividade exclusiva das grandes empresas de tecnologia, como as que exploram as mais populares redes sociais. Também moderam conteúdos de seus usuários (ou de terceiros) as enciclopédias online na curadoria dos verbetes e suas edições, as plataformas de reclamações e avaliações sobre produtos e serviços, os veículos de imprensa pelos comentários em notícias publicadas seus sites, além das empresas de *marketplace* na gestão dos anúncios que são postados pelos vendedores. Não raramente ferramentas de consultas públicas, organizadas por entidades governamentais, também demandam atividades de moderação para assegurar a produtividade do debate.

Com a ascensão dos agentes pessoais, *chats* e demais plataformas de criação de conteúdo, seja em texto, vídeo ou voz, a moderação de conteúdo também passa a ser um expediente adotado nesses espaços online, de modo a evitar que as respostas dadas por aplicações de inteligência artificial possam causar danos.[1]

Adicionalmente, existem empresas dedicadas à prestação de serviços de moderação de conteúdo, atuando como terceirizadas ou contratadas de outras empresas. Esse ecossistema é geralmente pouco conhecido ou estudado, adicionando mais complexidade no debate público sobre dinâmicas de moderação.[2]

1. KAMINSKI, Margot; JONES, Meg. ***Constructing AI Speech. Yale Law Journal Forum*** (22.04.2024), University of Colorado Law Legal Studies Research Paper No. 24-11. Disponível em: https://ssrn.com/abstract=4764706.

2. Vide ROBERTS, Sarah T. ***Behind the Screen: content moderation in the shadows of social media.*** New Haven: Yale University Press, 2019.

Por que isso importa?

Uma política pública ou projeto de lei que enderece o tema da moderação de conteúdo deve saber diferenciar o impacto, os recursos e as condições nas quais ocorre a moderação a partir das diferentes figuras envolvidas. Regras pensadas apenas para as grandes empresas de tecnologia podem não fazer sentido ou onerar demasiadamente a operação de pequenas empresas e entidades do terceiro setor.

02.

QUAL A NATUREZA DAS REGRAS SOBRE MODERAÇÃO DE CONTEÚDO?

Moderação de conteúdo é um tema transversal ao Direito Privado e ao Direito Público.

A definição de regras para ordenar a comunicação em ambientes digitais é um exercício da autonomia privada de empresas e demais entidades responsáveis por esses espaços. Os seus usuários, quando aceitam essas regras, ingressam em uma relação contratual.

Institutos característicos do Direito Privado, como o contrato e a responsabilidade civil, vêm sendo transformados na última década no sentido de lhes atribuir, para além de sua estrutura, uma função social atrelada ao atendimento de interesses públicos.[3] Fala-se também de eficácia horizontal dos direitos fundamentais nas relações privadas.[4]

O debate sobre moderação de conteúdo é um exemplo dessa transformação, já que as regras criadas para gerir um espaço privado passam a atrair o escrutínio sobre o seu impacto no exercício de direitos fundamentais, como a liberdade de expressão.[5] Essa é a porta de entrada para que autoridades públicas passem a acompanhar as práticas de moderação e considerá-las na formulação de políticas públicas e no exercício de atividades investigativas, por exemplo.

Por que isso importa?

Para além de uma questão teórica, reconhecer a característica de transversalidade entre o privado e o público das regras sobre moderação de conteúdo abre espaço para o debate sobre os contornos da autonomia privada, e revela os temas sobre os quais governos e demais autoridades precisam se debruçar para a proteção de direitos fundamentais e a repressão de práticas ilícitas.

3. TEPEDINO, Gustavo. "O princípio da função social no direito civil contemporâneo". ***In: Revista do Ministério Público do Estado do Rio de Janeiro, Rio de Janeiro***, nº 54 (out/dez 2014), p.141-154.
4. SARMENTO, Daniel. ***Direitos fundamentais e relações privadas.*** Rio de Janeiro: Lumen Juris, 2004.
5. MENDES, Gilmar Ferreira; FERNANDES, Victor Fernandes. "Eficácia dos direitos fundamentais nas relações privadas da internet: o dilema da moderação de conteúdo em redes sociais na perspectiva comparada Brasil-Alemanha". ***In: Revista de Direito Civil Contemporâneo***, v. 31, n. 9 (2022), p.33–68.

03.
MODERAÇÃO COMO PROCESSO

Moderação não se resume à decisão sobre um conteúdo, ela envolve atividades prévias e posteriores que devem ser observadas.

Estamos acostumados a pensar em moderação de conteúdo como a decisão sobre a remoção ou a manutenção de uma publicação em rede social. A literatura acadêmica, as propostas elaboradas pelo terceiro setor e diversas entidades[6], assim como as primeiras regulações surgidas no exterior, oferecem um panorama amplo sobre o tema, podendo assim ser proposta uma perspectiva da moderação como um verdadeiro processo.

Esse processo possui uma fase preparatória (anterior à tomada de decisão), a decisão em si, além de uma fase posterior, que permite avaliar as decisões tomadas. Cada uma dessas fases possui elementos próprios que, uma vez reunidos, compõem um sistema de moderação que viabiliza um olhar mais geral sobre as práticas adotadas pelas empresas ou pelas entidades responsáveis pela governança de ambientes online.

De modo sintético, pode-se assim visualizar as fases do processo de moderação e seus elementos constituintes:

6. Vide, dentre outros: 1) Princípios de Manila sobre Responsabilidade de Intermediários, disponível em: https://manilaprinciples.org/pt-br.html; 2) Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo, disponível em: https://santaclaraprinciples.org; 3) ***Change the Terms,*** disponível em: https://www.changetheterms.org; 4) ***Christchurch Call*** pela eliminação do terrorismo e do conteúdo extremista online, disponível em: https://www.christchurchcall.com/index.html; 5) ***Paris Call*** por Confiança e Segurança no Ciberespaço, disponível em: https://pariscall.international/en/principles.

MEDIDAS PRÉVIAS À TOMADA DE DECISÃO

1.

Conhecimento prévio das regras que regem o ambiente online (transparência);

2.

Criação de um canal de notificações (denúncias) sobre conteúdos lesivos que seja intuitivo e de fácil acesso;

3.

Treinamento permanente da equipe de moderadores;

4.

Aperfeiçoamento das ferramentas de moderação automatizadas.

MEDIDAS RELACIONADAS À TOMADA DE DECISÃO:

5.

Indicação clara sobre qual conteúdo infringente gerou a moderação;

6.

Indicação sobre qual política ou regra foi violada, de modo a evitar justificativas genéricas e que não possam municiar o usuário de fundamentos para recorrer da decisão.[7]

7. Os elementos 5 e 6 podem ser flexibilizados caso a moderação tenha ocorrido, por exemplo, no contexto de uma investigação conduzida por autoridades públicas.

MEDIDAS POSTERIORES À TOMADA DE DECISÃO:			
7.	**8.**	**9.**	**10.**
Criação de ferramenta que permita ao usuário recorrer da decisão de moderação;	Estabelecimento de política de salvaguardas que permita preservar o conteúdo removido para eventual restabelecimento ou produção de prova;	Elaboração de relatórios de transparência informativos e atualizados, endereçando a quantidade (números) e a qualidade da moderação, viabilizando assim análises sobre pontos fortes e oportunidades de melhoria;	Desenvolvimento de entidades de suporte (como conselhos consultivos) ou de supervisão para auxiliar na avaliação ou providenciar a revisão de casos críticos.

Por que isso importa?

A percepção das atividades de moderação como um processo facilita a compreensão sobre os diferentes elementos que devem ser observados para a formação de uma rotina de moderação robusta e equilibrada. De certa maneira, ela oferece um *checklist* sobre boas práticas, ferramentas e formas de organização que podem ser manejadas na governança de ambientes online.

04.

NÃO É SOBRE AS ÁRVORES, MAS SOBRE A FLORESTA

A moderação de conteúdo não deve ser avaliada a partir do erro ou do acerto de uma decisão específica, mas a partir do funcionamento do sistema como um todo.

Aqui seguem algumas métricas sobre moderação de conteúdo em grandes plataformas digitais entre os meses de julho e setembro de 2024. Nesse período, o YouTube removeu mais de 9 (nove) milhões de vídeos.8 A Meta estima que cerca de 3% (três por cento) dos usuários ativos mensais no Facebook são contas falsas, o que gerou mais de 1 (um) bilhão de contas moderadas entre os meses indicados.9 De acordo com o relatório de transparência do TikTok, só no Brasil, foram removidos 5.325.026 (cinco milhões, trezentos e vinte e cinco mil e seis) vídeos no período, sendo 89,7% deles em menos de 24 horas.10

É impossível acertar 100% das vezes quando a moderação recai sobre um volume gigantesco de conteúdo. A busca pelo aperfeiçoamento das atividades de moderação e a sua consequente supervisão passa pelo entendimento de que a avaliação, tanto interna das empresas, quanto externa por autoridades competentes, deve focar cada vez mais na organização dos diversos elementos que constituem o processo de moderação, e menos no erro ou acerto de uma decisão em si.[11]

Essa transição do foco sobre a decisão relativa a um conteúdo particular para a visão mais geral sobre o quanto a empresa está preparada para exercer as atividades de moderação impacta a maneira pela qual autoridades desenham políticas públicas sobre o tema, e pode até mesmo transformar a dinâmica da responsabilização das empresas e demais entidades.

---

8. Google. ***Cumprimento das Diretrizes da Comunidade do YouTube*** (julho-setembro/2024). Disponível em: https://transparencyreport.google.com/youtube-policy/removals.
9. Meta. ***Community Standards Enforcement Report*** (julho-setembro/2024). Disponível em: https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/.
10. TikTok. ***Community Guidelines Enforcement Report*** (julho-setembro/2024). Disponível em: https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2024-3.
11. DOUEK, Evelyn. "Content Moderation as Systems Thinking" (10.01.2022). ***In: Harvard Law Review Vol. 136.*** Disponível em: https://ssrn.com/abstract=4005326.

Existe ainda uma peculiaridade na maneira pela qual atividades de moderação são rotineiramente reportadas na imprensa: a depender do tema, do contexto ou dos usuários envolvidos, uma decisão errada ganha a repercussão que milhares de decisões acertadas não receberiam. A falta de transparência sobre as práticas de moderação também auxilia a fomentar suspeitas sobre os motivos e a extensão das atividades de governança de ambientes online.

Por que isso importa?

Mudar o foco para uma visão geral sobre as práticas adotadas por empresas e demais entidades permite não apenas o desenvolvimento de um debate mais informado sobre moderação, como também se permite o desenho de políticas públicas que não estejam aprisionadas na percepção de uma ou outra decisão equivocada que tenha obtido grande repercussão.

Decisões de moderação desproporcionais ou equivocadas em algum momento vão acontecer. Precisamos saber se as empresas estão comprometidas em preveni-las, de modo a reduzir a sua incidência, evidenciando a preparação e a execução de um sistema robusto de moderação de conteúdo para proteger os seus usuários e evitar a disseminação de conteúdos e comportamentos ilícitos em seus ambientes online. Para isso, vale lembrar, é preciso que as empresas também colaborem ao fornecer os dados que possam municiar essa análise.[12] A opacidade sobre moderação de conteúdo dificulta o debate e reduz a confiança da sociedade, podendo levar à edição de políticas públicas inadequadas sobre o tema.

Esse olhar mais geral, que por vezes aparece como a aplicação de um chamado "dever de cuidado", figura nas discussões travadas no Supremo Tribunal Federal por conta do julgamento de duas ações sobre o regime de responsabilidade civil dos provedores de aplicações, nas quais está sendo analisada a constitucionalidade do artigo 19 do Marco Civil da Internet.[13]

---

12. KELLER, Daphne; e LEERSSEN, Paddy. "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" In: N. Persily e J. Tucker, ***Social Media and Democracy: The State of the Field and Prospects for Reform*** (Cambridge University Press, 2020). Disponível em: https://ssrn.com/abstract=3504930.

13. SOUZA, Carlos Affonso. "STF julga ações que podem mudar a responsabilidade de plataformas no Brasil". In: UOL (27.11.2024), disponível em: https://www.uol.com.br/tilt/colunas/carlos-affonso-de-souza/2024/11/27/stf-julga-acoes-que-podem-mudar-a-responsabilidade-de-plataformas-no-brasil.htm.

05.
MODERAÇÃO É UM PRODUTO DO TEMPO E DO ESPAÇO

As decisões sobre moderação variam de acordo com a percepção social sobre os temas ao longo do tempo, assim como são determinadas pela cultura (e pelas leis) do país em que o conteúdo é avaliado.

Percepções sociais sobre os mais diversos temas variam no curso do tempo, impactando as atividades de moderação. Aqui reside um desafio que usualmente aparece nas falas de diretores de grandes empresas de tecnologia no sentido de que eles não querem ser os "árbitros da verdade"[14]. Se por um lado esse papel seria de todo indesejado, por outro é preciso reconhecer que uma parte relevante do debate público passa cada vez mais pelas plataformas digitais. Sendo assim, conhecer as regras que governam esses espaços e como elas são aplicadas se torna uma questão central.

Em seu vídeo anunciando mudanças nas regras sobre moderação nas plataformas da Meta, Mark Zuckerberg, CEO da empresa, afirmou que a eleição de Donald Trump nos Estados Unidos representaria um "ponto de inflexão cultural" que favoreceria a permanência de mais conteúdos nas plataformas, movendo o foco da moderação para postagens que representem ofensas graves.[15]

Moderação não se transforma apenas ao longo do tempo, mas também varia de acordo com o espaço em que é aplicada. Nesse sentido, é preciso que as atividades de moderação sejam adequadas às leis nacionais, ao mesmo tempo em que respeitam os direitos humanos reconhecidos internacionalmente. Esse equilíbrio representa uma tensão permanente, já que a tendência das empresas que atuam globalmente é se ater ao *compliance* de uma política e adotar práticas uniformes nos diferentes países em que atuam.

---

14. The Guardian. ***"Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat"*** (28.05.2020). Disponível em: https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump.

15. Meta. ***"More speech, fewer mistakes"***. Disponível em: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

Por que isso importa?

Moderação de conteúdo não é uma atividade monolítica. Ela acompanha as mudanças culturais e deve buscar atender às regras traçadas globalmente, mas sempre de olho nas peculiaridades das legislações nacionais, que podem fazer com que certas políticas possam receber mais atenção ou aplicações distintas a depender do contexto.

Conteúdos que são permitidos no Estados Unidos, dada a trajetória jurídica do país na construção da liberdade de expressão, podem ser reputados ilícitos em outras jurisdições, demandando assim uma adequação das práticas de moderação.

06.

MODERAÇÃO É AFETADA POR TENDÊNCIAS TECNOLÓGICAS

A percepção social sobre o que tecnologias emergentes podem fazer, como aplicações de inteligência artificial, moldam o debate sobre moderação de conteúdo. A tendência em descentralizar a moderação (por meio de notas da comunidade) precisa de mais estudos para que possa ser devidamente avaliada.

Desde a introdução do ChatGPT, no final de 2022, os debates sobre o que aplicações de inteligência artificial podem ou não fazer tomaram o imaginário popular. No senso comum, a impressão de que ferramentas de IA superam em muito as capacidades humanas de processamento de informações faz com que um "efeito de substituição"[16] se manifeste nas mais diversas atividades, inclusive na moderação de conteúdo.

Como lembra Evelyn Douek: "existem dois tipos de ferramentas automatizadas usadas na moderação de conteúdo comercial: sistemas de correspondência, que comparam novas postagens com um banco de dados de conteúdos previamente classificados, e sistemas preditivos, que buscam classificar novas postagens como contrárias às regras da plataforma."[17]

Ambos os modelos são fundamentais para dar escala à moderação de conteúdo, mas ambos também erram. Sistemas de correspondência podem ter falsos positivos e negativos (marcando como inapropriados conteúdos lícitos ou falhando em reconhecer um conteúdo já sinalizado como ilícito por causa de pequenas modificações na foto ou vídeo original, por exemplo). Sistemas preditivos usualmente carecem de maior transparência, sendo o seu desenvolvimento, operação e avaliações pouco conhecidos.[18]

16. BALKIN, Jack M. ***"The Path of Robotics Law". In: California Law Review, v.6*** (junho/2015). Yale Law School, Public Law Research Paper No. 536. Disponível em: https://ssrn.com/abstract=2586570.

17. DOUEK, Evelyn. ***"Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability".*** 121 COLUM. L. REV. 759 (2021). Disponível em: https://ssrn.com/abstract=3679607.

18. GORWA, R.; BINNS, R.; e KATZENBACH, C. ***"Algorithmic content moderation: Technical and political challenges in the automation of platform governance".*** In: Big Data & Society, 7 (2020). Disponível em: https://doi.org/10.1177/2053951719897945.

Mais recentemente, algumas empresas de redes sociais vêm adotando um modelo descentralizado de moderação, rotineiramente chamado de “notas da comunidade”, no qual se permite que os próprios usuários das plataformas possam agregar contexto ou mesmo refutar o conteúdo de uma publicação.

A produção das notas da comunidade passa por um sistema de avaliação entre os usuários, tornando o conteúdo da nota visível para todos na plataforma caso o seu conteúdo alcance um determinado nível de consenso sobre a sua fatualidade, pertinência e utilidade.

Essa prática, que deve complementar as atividades de moderação desenvolvidas pelas empresas, vem sendo apresentada como solução para alguns dos dilemas na governança de ambientes online.

A literatura acadêmica e especializada sobre o tema ainda é incipiente e apenas analisa a experiência inicial de implementação de notas da comunidade pela rede social X (antigo Twitter). De toda forma, reunimos aqui alguns dos principais achados das publicações sobre o tema como forma de orientar a análise dessa questão pelas autoridades públicas e demais interessados.

PONTOS POSITIVOS DAS NOTAS DA COMUNIDADE:

1. Probabilidade de que postagens sejam deletadas.

Publicações que recebem notas da comunidade visíveis ao público têm maior probabilidade de serem deletadas pelo autor do conteúdo original.[19] Um estudo sugere que a chance de a publicação ser deletada é de 80% (oitenta por cento), reduzindo pela metade a chance de o conteúdo ser retuitado.[20]

2. Mais contexto.

O contexto fornecido pelas notas da comunidade aumenta a confiança do público em comparação com as ferramentas de rotulagem que apenas sinalizam que publicações contêm informações falsas ou enviesadas.[21]

19. GAO, Yang; ZHANG, Maggie Mengqing; e RUI, Huaxia. ***"Can Crowdchecking Curb Misinformation? Evidence from Community Notes"***. Disponível em https://ssrn.com/abstract=4992470.
20. RENAULT, Thomas; RESTREPO-AMARILES, David; e TROUSSEL-CLÉMENT, Aurore. ***"Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News"***. Disponível em https://ssrn.com/abstract=4800565.
21. DROLSBACH, Chiara Patricia; SOLOLEV, Kirill; e PRÖLLOCHS, Nicolas. ***"Community Notes Increase Trust in Fact-Checking on Social Media"***. PNAS Nexus, v. 3, 2024. Disponível em: https://doi.org/10.1093/pnasnexus/pgae217.

PONTOS NEGATIVOS DAS NOTAS DA COMUNIDADE:

1. Famosos na mira. Contribuidores das notas da comunidade priorizam verificar postagens de contas influentes com muitos seguidores, possivelmente para maximizar o impacto da moderação.[22]	**2. Pouca diversidade.** O sistema enfrenta problemas de representatividade, já que poucos contribuidores monopolizam a criação de notas da comunidade, facilitando situações de captura e amplificação de certos pontos de vista.[23]	**3. Lentidão.** O tempo necessário para criar e aprovar notas torna o sistema ineficaz para lidar com conteúdos virais em suas fases iniciais.[24]

22. PILARSKI, Moritz; SOLOVLEV, Kirill Olegovich; e PRÖLLOCHS, Nicolas. ***"Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media"***. In: Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM), 2024. Disponível em: https://arxiv.org/abs/2305.09519.

23. WIRTSCHAFTER, Valerie; e MAJUMDER, Sharanya. ***"Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter's Community Notes".*** In: Journal of Online Trust and Safety, Stanford Internet Observatory, v. 2, n. 1, 2023. Disponível em: https://www.tsjournal.org/index.php/jots/article/view/139.

24. CHUAI, Yuwei; TIAN, Haoye; PRÖLLOCHS, Nicolas; e LENZINI, Gabriele. ***"Did the Roll-Out of Community Notes Reduce Engagement with Misinformation on X/Twitter?"*** In: Proceedings of the ACM Human-Computer Interaction, v. 8, CSCW2, 2024. Disponível em: https://dl.acm.org/doi/10.1145/3686967.

4. Quantidade x Qualidade.

A estrutura atual do sistema privilegia métricas quantitativas (p.ex. número de interações) em detrimento de análises qualitativas (como a gravidade dos danos causados pelo conteúdo).[25]

5. Polarização e consenso.

Assuntos controvertidos e polarizados são especialmente problemáticos no período eleitoral, pois raramente atingem consenso entre os revisores para se aprovar a nota da comunidade e torná-la visível. Um estudo revela que 74% das notas que tratavam de desinformação eleitoral, e que citavam fontes confiáveis, não foram exibidas por falta de consenso.[26] Outro aponta que no Brasil, como um todo, apenas 8% das notas criadas obtêm consenso e chegam ao público.[27]

Por que isso importa?

Por ser tão permeável às tendências tecnológicas, a compreensão sobre as atividades de moderação de conteúdo precisa vir sempre acompanhada de uma revisão da literatura acadêmica e especializada, além dos conhecimentos técnicos mais atualizados, de modo a evitar a reprodução de lugares-comuns que podem enviesar uma política pública ou tornar ineficaz uma tomada de decisão pela autoridade competente.

Especialmente no que diz respeito ao uso de ferramentas de inteligência artificial, é preciso estar alerta para as possibilidades e limitações desses recursos. Sobre as chamadas "notas da comunidade", é preciso avançar na pesquisa sobre o desenho de sistemas de moderação de conteúdo descentralizados e seus resultados – positivos ou negativos – em redes sociais de larga escala.

---

25. MATAMOROS FERNÁNDEZ, Ariadna; e JUDE, Nadia Alana. ***"The Importance of Centering Harm in Data Infrastructures for 'Soft Moderation': X's Community Notes as a Case Study"***. In: New Media and Society. Special issue Infrastructures for Datafication, edited by Stine Lomborg, Alessandro Gandini, Jennifer Pybus and Signe Sophus Lai. DOI: 10.1177/14614448251314399/.

26. Center for Countering Digital Hate (CCDH). Rated Not Helpful: ***How X's Community Notes System Falls Short on Misleading Election Claims***. CCDH, 2024. Disponível em: https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf.

27. TARDÁGUILA, Cristina. ***"Only 8% of Community Notes in Portuguese on X reach users"***. Lupa. Disponível em:https://lupa.uol.com.br/jornalismo/2025/01/09/only-8-of-community-notes-in-portuguese-on-x-reach-users.

07.

O PAPEL DOS CONSELHOS DE SUPERVISÃO.

A experiência acumulada de uma empresa com moderação de conteúdo pode criar uma visão ensimesmada de suas regras e procedimentos. Conselhos consultivos e de supervisão ajudam a quebrar esse ciclo.

Diferentes empresas criaram, nos últimos anos, conselhos consultivos e de supervisão com relação às atividades de moderação de conteúdo.[28] Esses conselhos podem servir como uma forma de institucionalizar o contato com especialistas da academia e do terceiro setor, oferecendo um espaço para análise crítica das práticas de moderação adotadas pelas empresas.

Ao mesmo tempo, conselhos de supervisão, cujo exemplo mais conhecido é o Comitê de Supervisão da Meta[29], têm por objetivo analisar e revisar decisões tomadas em casos concretos, sugerindo ainda aperfeiçoamentos nas políticas das empresas.

Um efeito relevante da figura do conselho de supervisão é colocar as tomadas de decisão sobre moderação de conteúdo sob a perspectiva dos compromissos que os países e as empresas têm através dos Princípios Orientadores das Nações Unidas sobre Empresas e Direitos Humanos[30].

Com isso, conselhos de supervisão podem estabelecer um suporte objetivo para as decisões sobre moderação de conteúdo, amparadas pelas décadas de experiência na aplicação de decisões sobre liberdade de expressão no contexto internacional dos direitos humanos.

---

28. Vide, por exemplo, o ***Spotify Safety Advisory Council*** (https://newsroom.spotify.com/2022-06-13/introducing-the-spotify-safety-advisory-council/) e o ***Safety and Content Advisory Council*** do TikTok (https://www.tiktok.com/transparency/en-us/advisory-councils/).

29. Comitê de Supervisão da Meta (https://www.oversightboard.com/?lang=pt-br). Para comentários e críticas sobre a figura do comitê de supervisão, vide ESTARQUE, Marina; ARCHEGAS, João Victor. ***Redes Sociais e Moderação de Conteúdo: criando regras para o debate público a partir da esfera privada. Rio de Janeiro: Instituto de Tecnologia e Sociedade,*** 2021. Disponível em: https://itsrio.org/pt/publicacoes/redes-sociais-e-moderacao-de-conteudo; e MANTUANI, Matheus. ***"Desafios à Moderação de Conteúdo no Facebook"***. In: Sergio Branco e Chiara de Teffé (org.) ***Regulação Digital: perspectivas jurídicas sobre tecnologia e sociedade. Rio de Janeiro: Instituto de Tecnologia e Sociedade***, 2024. Disponível em: https://itsrio.org/wp-content/uploads/2016/12/20240119_Livro_Pos_6_ITS-UERJ_COMPLETO.pdf

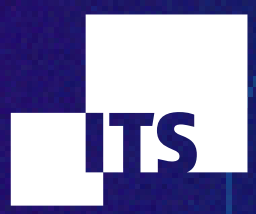
30. Organização das Nações Unidas. ***Guiding Principles on Business & Human Rights***. Disponível em: https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

Esse papel é especialmente relevante no contexto em que as regras criadas pelas empresas para sua operação global podem acabar gerando uma cultura tão autorreferenciada de moderação que, no final das contas, termina por perder conexão com os padrões adotados em documentos internacionais e legislações nacionais ao redor do mundo.[31]

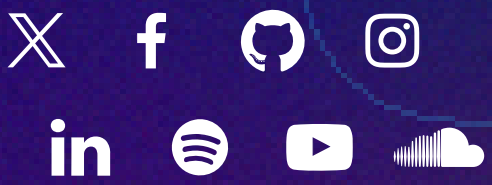
Por que isso importa?

Essa experiência pode ajudar a informar e balizar o debate sobre moderação de conteúdo, evitando que ele se torne casuístico e definido pelo calor do momento, tornando-se, em vez disso, pensado a partir de uma visão de longo prazo, com base em normas de validade internacional que são obrigatórias para os países que as subscreveram, incluindo os Estados Unidos e o Brasil.

31. KETTEMANN, Matthias C.; SCHULZ, Wolfgang: ***"Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study".*** Hamburg: Working Papers of the Hans-Bredow-Institut # 1 (Janeiro/2020). Disponível em: https://doi.org/10.21241/ssoar.71724



Acesse nossas redes



itsrio.org