Pontifícia Universidade Católica
do Rio de Janeiro

Juliana Libman

MODERAÇÃO DE CONTEÚDO EM REDES SOCIAIS: POR UMA REGULAÇÃO QUE PROMOVA A LIBERDADE DE EXPRESSÃO

Dissertação de Mestrado

Dissertação apresentada à banca examinadora como requisito parcial para obtenção do grau de Mestre pelo Programa de Mestrado Profissional em Direito Civil Contemporâneo e Prática Jurídica da Pontifícia Universidade Católica (PUC-Rio).

Prof. Marcelo Calixto

Orientador

Departamento de Direito - PUC-Rio

Prof.ª Chiara de Teffé

Coorientadora

Departamento de Direito - UERJ

Rio de Janeiro
Março de 2023

PUC-Rio - Certificação Digital Nº 2112697/CA

Pontifícia Universidade Católica
do Rio de Janeiro

Juliana Libman

MODERAÇÃO DE CONTEÚDO EM REDES SOCIAIS: POR UMA REGULAÇÃO QUE PROMOVA A LIBERDADE DE EXPRESSÃO

Dissertação apresentada à banca examinadora como requisito parcial para obtenção do grau de Mestre pelo Programa de Mestrado Profissional em Direito Civil Contemporâneo e Prática Jurídica da Pontifícia Universidade Católica (PUC-Rio).

Prof. Marcelo Calixto

Orientador

Departamento de Direito - PUC-Rio

Prof.ª Chiara de Teffé

Coorientadora

Departamento de Direito - UERJ

Prof.ª Ana Frazão
Professora de Direito Civil, Comercial e Econômico da Universidade de Brasília (UnB)

Prof. Carlos Affonso de Souza
Professor da PUC-Rio.

Rio de Janeiro, 13 de março de 2023.

PUC-Rio - Certificação Digital Nº 2112697/CA

Todos os direitos reservados. A reprodução, total ou parcial do trabalho, é proibida sem autorização da universidade, da autora e do orientador.

Juliana Libman

Mestre em Direito Civil pela PUC-Rio.

Ficha Catalográfica

Libman, Juliana

Moderação de conteúdo em redes sociais : por uma regulação que promova a liberdade de expressão / Juliana Libman ; orientador: Marcelo Calixto ; coorientadora: Chiara de Teffé. – 2023.

137 f. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Direito, 2023.

1. Direito – Teses. 2. Redes sociais. 3. Moderação de conteúdo. 4. Liberdade de expressão. 5. Regulação. 6. Autorregulação. I. Calixto, Marcelo. II. Teffé, Chiara de. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Direito. IV. Título.

CDD: 340

PUC-Rio - Certificação Digital Nº 2112697/CA

Agradecimentos

Agradeço a minha família, e especialmente aos meus pais, Aida e Elisio, por todo o carinho, força e dedicação, que nunca mediram esforços para me proporcionar a melhor forma de educação, e que sempre me incentivaram e apoiaram para que eu alcançasse meus sonhos. Junto a eles, agradeço aos meus avós, pelos exemplos de vida, perseverança e dedicação.

Agradeço ainda às minhas irmãs, Gabriela e Priscila, minhas primeiras e eternas companheiras de vida, que para sempre estarão comigo, e ao meu noivo, Eduardo, por todo suporte e parceria ao longo dessa jornada.

Da mesma maneira, às minhas queridas amizades, que com muito amor construí ao longo de minha vida. Aos amigos que tenho desde a infância, e aqueles que tive o imenso prazer de conhecer na faculdade.

A todos os professores, pela dedicação empenhada ao longo dos cinco anos de curso.

Agradeço, ainda, aos meus companheiros de trabalho, que muito me incentivaram e transformaram o ambiente de trabalho em local agradável, de companheirismo e admiração. Dessa forma, agradeço especialmente aos meus chefes, Marcela e Felipe, pela amizade e contribuição para o meu crescimento pessoal e profissional.

Por fim, agradeço aos meus orientadores neste trabalho, Marcelo Calixto e Chiara de Teffé, pela atenção, empenho, e atenção durante toda a orientação. Quem não mediram esforços para me ajudar em cada etapa deste estudo e, muito além disso, tornaram-se exemplos para mim de profissionais dedicado e apaixonados pelo que fazem, estudam e lecionam: o Direito.

PUC-Rio - Certificação Digital Nº 2112697/CA

Resumo

LIBMAN, Juliana. **Moderação de conteúdo em redes sociais: Por uma regulação que promova a liberdade de expressão**. Orientador: CALIXTO, Marcelo. Coorientadora: DE TEFFÉ, Chiara. Rio de Janeiro, 2023. 137 p. Dissertação de Mestrado – Departamento de Direito, Pontifícia Universidade Católica do Rio de Janeiro.

O presente estudo visa a apresentar uma análise acerca da atividade de moderação de conteúdo exercida pelas grandes redes sociais, como Facebook, Instagram e Twitter. Para tanto, inicialmente, serão analisados como as redes sociais alcançaram o status que têm hoje como grandes influenciadoras do discurso público e como a atividade de moderação de conteúdo se desenvolveu e se encontra na atualidade, sendo uma atividade necessária para assegurar o direito à liberdade de expressão dos usuários, que devem ser resguardos de espaços digitais tóxicos, com a proliferação de conteúdo ilícito e falso. Em seguida, serão apresentados os desafios envolvidos na atividade de moderação de conteúdo. Por fim, com base nos conceitos e problemas apresentados, será feita uma análise da melhor forma para se regular a atividade de moderação de conteúdo, em consonância com as normas do Marco Civil da Internet (Lei nº 12.965/2014).

PUC-Rio - Certificação Digital Nº 2112697/CA

Palavras-chave

Redes Sociais; Moderação de Conteúdo; Liberdade de Expressão; Regulação; Autorregulação.

Abstract

LIBMAN, Juliana. **Content moderation in social media: For a regulation that promotes the freedom of expression**. Orientador: CALIXTO, Marcelo. Coorientadora: DE TEFFÉ, Chiara. Rio de Janeiro, 2023. 137 p. Dissertação de Mestrado – Departamento de Direito, Pontifícia Universidade Católica do Rio de Janeiro.

The present study aims to present an analysis about the content moderation activity carried out by the major social networks, such as Facebook, Instagram and Twitter. To do so, initially, it will be analyzed how social networks reached the status they have today as major influencers of public discourse and how the content moderation activity has developed and is currently found, being a necessary activity to ensure the right to freedom of expression of users, who must be safeguarded of toxic digital spaces, with the proliferation of illicit and false content. Next, the challenges involved in the content moderation activity will be presented. Finally, based on the concepts and problems presented, an analysis will be made of the best way to regulate content moderation, in accordance with the Internet Legal Framework (Law No. 12,965/2014).

PUC-Rio - Certificação Digital Nº 2112697/CA

Keywords

Social Media; Content Moderation; Freedom of Expression; Regulation; Self-regulation.

Sumário

PUC-Rio - Certificação Digital Nº 2112697/CA

PUC-Rio - Certificação Digital Nº 2112697/CA

Lista de abreviações

CDA – *Communication Decency Act*

CDT – *Center for Democracy and Technology*

CGI.br – Comitê Gestor da Internet no Brasil

DCMA – *Digital Millenium Copyright Act*

DCE – Diretiva de Comércio Eletrônico

DSA – *Digital Services Act*

MP – Medida Provisória

MPF – Ministério Público Federal

NetzDG – *Network Enforcement Act*

OFCOM – *Office of Communication*

PL – Projeto de Lei

STF – Supremo Tribunal Federal

STJ – Superior Tribunal de Justiça

SL – Suspensão Liminar

TJMG – Tribunal de Justiça de Minas Gerais

TJSC – Tribunal de Justiça de Santa Catarina

TJSP – Tribunal de Justiça de São Paulo

PUC-Rio - Certificação Digital Nº 2112697/CA

INTRODUÇÃO

Redes sociais podem ser definidas como "serviços materializados em páginas na Web ou em aplicativos que, a partir de perfis pessoais, permitem uma ampla interação entre seus usuários, proporcionando e facilitando as relações e os laços sociais entre os sujeitos[1]" no ambiente virtual. Como observam Maria Celina Bodin de Moraes e Chiara de Teffé, as redes sociais têm as seguintes características:

> i) a existência de um ambiente propício à interação entre os usuários na plataforma; ii) o pedido de dados pessoais para a criação de perfis, que são vinculados a contas determinadas; iii) a articulação de uma lista de outros usuários com os quais se compartilha conexões; e iv) o oferecimento de ferramentas que permitem e estimulam que o usuário adicione seu próprio conteúdo na rede, como fotografias, comentários, músicas, vídeos ou links para outros sites, de modo que ocorra a expansão da estrutura da própria rede social[2].

A ascensão das grandes redes sociais[3] teve impacto significativo na liberdade de expressão. Elas permitiram que as pessoas compartilhassem suas opiniões e ideias de maneira fácil e acessível, ampliando o acesso à informação e ao debate público. Isso também possibilitou grupos marginalizados e minorias de terem uma voz mais forte e serem ouvidos em questões importantes. Com efeito, o advento das plataformas de redes sociais como novo meio de comunicação se mostrou "importante variável na estrutura democrática dos países, seja em nível global, seja em nível local"[4].

Assim, se antes a participação no debate público estava sujeita ao controle editorial exercido sobre veículos de mídias tradicionais, centralizado em jornais, rádios e canais de televisão, com o exercício do direito à liberdade de expressão

PUC-Rio - Certificação Digital Nº 2112697/CA

[1] MORAES, Maria Celina Bodin de; TEFFÉ, Chiara Spadaccini de. *Redes sociais virtuais: privacidade e responsabilidade civil. Análise a partir do Marco Civil da Internet*. Pensar, Fortaleza, v. 22, n. 1, p. 108-146, jan./abr. 2017, p. 10. Disponível em: <https://ojs.unifor.br/rpen/article/view/6272>. Acesso em: 10 nov. 2022.
[2] Ibid.
[3] O presente estudo se concentra na análise de grandes redes sociais como Facebook, Instagram, Twitter, TikTok e YouTube. Há, contudo, uma priorização de análise de casos do Facebook, por ser a plataforma mais utilizada e a que mais disponibilizou informações públicas sobre moderação de conteúdo. Devido à atualidade do tema e constantes atualizações na matéria, informa-se que a pesquisa que embasou o presente estudo foi concluída em novembro de 2022.
[4] FONSECA, Gabriel; VERONESE, Alexandre. Desinformação, fake news e mercado único digital: a potencial convergência das políticas públicas da União Europeia com os Estados Unidos para melhoria dos conteúdos comunicacionais. *In*: *Fake News e as eleições 2018*. Cadernos Adenauer, 2018, p. 43.

concentrado entre o Estado, de um lado, e essas organizações de mídia, de outro[5], com o advento das redes sociais isso mudou.

As redes sociais e o advento de novas tecnologias acabaram por encerrar a dependência havida nos veículos de mídias tradicionais. As redes sociais se tornaram as novas praças públicas, possibilitaram a criação de um espaço para debate público de forma *online*, sem controle editorial prévio da forma como ocorre com os veículos de mídias tradicionais. O discurso saiu da centralização das grandes mídias tradicionais, para descentralização do espaço digital.

No entanto, o crescimento das redes sociais e seu uso por bilhares de pessoas ao redor do mundo permitiram a apropriação dessas comunidades abertas para uso abusivo, com ampla disseminação de discursos de ódio, notícias falsas, e conteúdo ilícito ou ilegal. Como não podia deixar de ser, tornou-se necessário que as plataformas de redes sociais moderassem, de alguma maneira, o conteúdo nelas postado por usuários. Assim, as plataformas de redes sociais, que no início sugiram como meras empresas de tecnologia, passaram a exercer influência e controle sobre o discurso, tornando-se, para alguns autores, verdadeiras governantes[6] de espaços digitais.

As plataformas se viram, então, diante do desafio de poder influenciar o que pode ou não acontecer no espaço digital, o que tomou a forma de moderação de conteúdo. Na medida em que as plataformas ganharam popularidade e passaram a impor restrições sobre o discurso *online*, aumentou também a pressão de entidades externas sobre as decisões de moderação de conteúdo das plataformas.

Em 2011, por exemplo, um professor francês propôs ação contra o Facebook pelo cancelamento de sua conta "sem aviso ou justificativa", após a postagem de uma foto do quadro "A origem do mundo", de Gustave Courbert, que retrata uma vagina[7]. Nesse momento, havia o senso comum de que em redes sociais era

PUC-Rio - Certificação Digital Nº 2112697/CA

[5] BALKIN, Jack M. *Free Speech is a Triangle*. Columbia Law Review, v. 118, n. 07, p. 2011/2056, 2018. Disponível em: <https://columbialawreview.org/content/free-speech-is-a-triangle/>. Acesso em 12 out. 2022.

[6] O Facebook, maior plataforma existente hoje, por exemplo, conta com 2.9 bilhões de usuários mensais ativos, o que representa mais de um quarto da população mundial. Disponível em: <https://www.oficinadanet.com.br/post/16064-quais-sao-as-dez-maiores-redes-sociais>. Acesso em: 20 nov. 2022.

[7] Disponível em: <https://exame.com/tecnologia/facebook-exclui-usuario-que-postou-a-origem-do-mundo-de-courbet/>. Acesso em: 20 nov. 2022.

permitido publicar livremente, pois as regras de permissão ou proibição de discursos nas plataformas eram opacas[8].

Apenas em 2017 o Facebook passou a publicizar sua política de moderação de conteúdo. Uma reportagem do jornal britânico *The Guardian* trouxe um panorama inicial sobre o sistema de moderação de conteúdo da empresa e seu alcance, que incorporava controvérsias morais, políticas e jurídicas[9]. A reportagem deixou claro que o Facebook buscou traçar regras, por meio de sua política de moderação de conteúdo, de aplicação global, para identificar os limites ao "legítimo" exercício da liberdade de expressão na plataforma.

Surgiram, então, válidos questionamentos sobre o receio de censura e de remoções injustificadas de conteúdo por empresas privadas, especialmente devido à falta de regulamentação sobre a matéria. Há quem defenda que as plataformas interferem de forma exagerada no discurso público, limitando a liberdade de expressão no espaço *online*. Há quem, em sentido oposto, entenda que as plataformas interferem pouco, permitindo a existência de um ambiente digital tóxico, com a presença conteúdo indesejável.

De fato, a atividade de moderação de conteúdo não é uma tarefa simples, e apresenta diversos desafios. Moderar conteúdo é difícil diante do volume global de conteúdo postado por dia nas plataformas e da necessidade de tomada de decisões rápidas, diante do risco de viralização, por exemplo. Isso requer o uso de ferramentas automatizadas que, embora necessárias, apresentam seus próprios riscos e limitações, pois a inteligência artificial muitas vezes não está apta a compreender o contexto e o real sentido de uma postagem. Além disso, fato é que algoritmos que atuam por meio do *machine learning* incorporam vieses de quem cria o código, que podem ser discriminatórios[10]. Há, ainda, a dificuldade na interpretação de determinado conteúdo ao se considerar que as plataformas atuam em nível global, para bilhões de usuários, de diferentes culturas e contextos.

Se a moderação de conteúdo em si não é simples, a regulação dessa atividade também não é. A ausência de regulação acabou permitindo que grandes empresas

[8] Em 2015, os termos de uso do Facebook deixaram expresso que retratos de nudez em obras de arte eram permitidos. Em 2019 o processo chegou ao fim, com acordo entre as partes.
[9] Disponível em: <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>. Acesso em: 20 nov. 2022.
[10] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais. O problema jurídico da remoção de conteúdo pelas plataformas*. Belo Horizonte: Editora Dialética, 2021, p. 58.

PUC-Rio - Certificação Digital Nº 2112697/CA

privadas criassem regras sobre o discurso público de forma oculta e opaca, sem transparência ou devido processo legal aos usuários, na perspectiva da autorregulação.

Pensar em uma regulação estatal da internet, por sua vez, implica no risco de o Estado atuar com base em seus interesses próprios, censurando discursos que entenda prejudiciais, em violação ao direito fundamental à liberdade de expressão.

A regulação da moderação de conteúdo pode ser feita com base em três principais abordagens: (i) leis *antitruste*, que buscam incentivar a concorrência e criar os estímulos mercadológicos para que as plataformas ajam conforme os interesses de usuários; (ii) leis de proteção à privacidade, que garantem aos usuários maior controle sobre seus dados e/ou limitam o potencial de direcionamento de conteúdo pelas plataformas; e (iii) leis sobre responsabilização de intermediários pelo conteúdo postado por terceiros, que pretendem desenvolver um modelo de responsabilização civil que crie os incentivos adequados para que as plataformas promovam a liberdade de expressão, ao mesmo tempo em que combatam conteúdo danoso[11].

Sem desconsiderar a importância das duas primeiras abordagens, o objeto do presente trabalho está restrito ao terceiro plano. Assim, importa saber: o Marco Civil da Internet permite que redes sociais removam ou tornem indisponível conteúdo por decisão própria? Se sim, como regular a moderação de conteúdo pelas redes sociais? Quais os parâmetros normativos que devem nortear as condutas das redes sociais à luz da liberdade de expressão? Podem essas empresas privadas ser responsabilizadas por, ao aplicarem seus termos de uso, removerem determinado conteúdo das redes, que não seja ilegal?

Dessa forma, o presente trabalho tem por objetivo demonstrar os problemas e dificuldades enfrentados pelas plataformas de redes sociais na moderação de conteúdo postado pelos usuários, e a melhor forma de regular essa atividade. Isso significa pensar em uma forma que crie os incentivos adequados para a moderação de conteúdo na internet, protegendo manifestações legítimas de interferências

[11] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 38.

PUC-Rio - Certificação Digital Nº 2112697/CA

indevidas, e o debate público e outros usuários de ataques coordenados que buscam silenciá-los ou atacar instituições e o regime democrático[12].

PUC-Rio - Certificação Digital Nº 2112697/CA

[12] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 39.

Capítulo 1 – A moderação de conteúdo feita pelas redes sociais como forma de regulamentação da liberdade de expressão

Este primeiro capítulo apresenta os principais conceitos necessários para o desenvolvimento do estudo. Inicialmente, serão tecidas considerações acerca da liberdade de expressão e a concepção desse direito fundamental na atualidade, onde a maior parte do discurso público é exercido de forma *online*, por meio das grandes plataformas de redes sociais, como o Facebook, Instagram, Twitter, YouTube e TikTok.

Com efeito, se antes a liberdade de expressão era pensada para assegurar o discurso público em redes de mídias tradicionais, como rádio e televisão, em que se tem amplo controle editorial, essa perspectiva tradicional não se enquadra quando se pensa no discurso digital.

Nesse sentido, serão apresentadas, em linhas gerais, as formas pelas quais empresas privadas passaram a adotar um papel importante na construção e regulação dos novos paradigmas da liberdade de expressão no âmbito da internet, influenciando o discurso público. Isso se deu através do exercício da moderação do conteúdo postado por usuários em redes sociais, que, como será exposto, é fundamental para garantir a liberdade de expressão *online*.

Isso não significa que Estados e governos não participem mais do controle do discurso público, o que se dá através de leis e decisões judiciais, por exemplo. Mas será apresentado um panorama sobre como a internet e as grandes redes sociais propiciaram um ambiente de discurso público complexo, no qual a regulação desse discurso por parte dos Estados nacionais ocorre de forma conjunta com a regulação feita por atores privados e transnacionais.

1.1 A concepção de liberdade de expressão exercida nas plataformas de redes sociais e a moderação de conteúdo

PUC-Rio - Certificação Digital Nº 2112697/CA

A concepção tradicional da liberdade de expressão, bem explicada por Tim Wu[13] e William Marshall[14], foi pensada para um mundo em que a informação era escassa e "a participação no debate público dependia de investimentos financeiros elevados e de disputas de meios escassos, como frequências de rádio"[15]. Nessa concepção, predominava o entendimento de que a intervenção do Estado sobre o discurso seria uma ameaça à autonomia e à democracia. A liberdade era vista, assim, como uma liberdade negativa, "que impunha ao Estado um dever de abstenção, sob o fundamento de que atribuir a instituições políticas o poder para decidir o que pode ou não ser dito é perigoso, arbitrário e ilegítimo"[16].

Com o advento das redes sociais[17] e das novas tecnologias, contudo, essa concepção mudou. Antes a participação no debate público estava sujeita ao controle editorial exercido sobre veículos de mídias tradicionais, centralizado em jornais, rádios e canais de televisão, com o exercício do direito à liberdade de expressão concentrado entre o Estado, de um lado, e essas organizações de mídia, de outro[18]. Essa concentração acabava por limitar a participação de cidadãos comuns no debate público, além de facilitar a censura estatal prévia[19].

As redes sociais e o advento de novas tecnologias acabaram por encerrar a dependência havida nos veículos de mídias tradicionais. As redes sociais se tornaram as novas praças públicas, possibilitando a criação de um espaço para debate público de forma *online*, sem controle editorial prévio. O discurso saiu, assim, da centralização dos grandes veículos de mídias tradicionais, para descentralização do espaço digital. Como observa Manuel Castells, as redes sociais são espaços de autonomia além do controle de governos e empresas que, ao longo

PUC-Rio - Certificação Digital Nº 2112697/CA

[13] WU, Tim. Is the First Amendment Obsolete? *In:* POZEN, David E. (Ed.). *The Perilous Public Square*, New York: Columbia University Press, E-book Kindle. (Não paginado).
[14] MARSHALL, William P. The Truth Justification for Freedom of Speech. *In:* STONE, Adrienne; SCHAUER, Frederick (EDs.). *Freedom of Speech*. United Kingdom: Oxford University Press, 2021, p. 44-60.
[15] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 45.
[16] Ibid, p. 46.
[17] "Rede social é gente, é interação, é troca social. É um grupo de pessoas, compreendido através de uma metáfora de estrutura, a estrutura de rede. Os nós da rede representam cada indivíduo e suas conexões, os laços sociais que compõem os grupos. Esses laços são ampliados, complexificados e modificados a cada nova pessoa que conhecemos e interagimos." RECUERO, Raquel. Rede social. *In:* AVORIO, A.; SPYER, J. (Org.). *Para entender a Internet*. Versão rev. e ampl., 2015, p. 83.
[18] BALKIN, Jack M. *Free Speech is a Triangle*. Columbia Law Review, v. 118, n. 07, p. 2011/2056, 2018. Disponível em: <https://columbialawreview.org/content/free-speech-is-a-triangle/>. Acesso em: 12 out. 2022.
[19] BARROSO, Luna van Brussel. op. cit., p. 92.

da história, monopolizavam os canais de comunicação como alicerces de seu poder[20]. Ainda, como observa Luna Barroso:

> De forma inovadora, as plataformas digitais criaram comunidades online, para compartilhamento de textos, imagens, vídeos e links produzidos pelos próprios usuários, sem controle editorial. Ofereciam, portanto, um espaço aberto para que qualquer usuário compartilhasse conteúdo, de forma pública ou privada, sem depender de recursos financeiros, de intermediação de veículos de mídia ou de aprovação por conselhos de redação ou editoriais. Ao fazê-lo, facilitaram o discurso, diversificaram as fontes e multiplicaram exponencialmente a quantidade de informação disponível. Em 2018, mais de 3 bilhões de pessoas usavam redes sociais e, atualmente, apenas o Facebook tem mais de 2.5 bilhões de usuários, que podem compartilhar informações sobre os mais diversos assuntos[21].

A superação da concentração do discurso público nas mídias tradicionais facilitou sua democratização, dando voz a minorias e diversificando o debate público, pois permitiu que qualquer indivíduo tenha voz para milhares de outros usuários[22]. Sobre o tema, Manuel Castells explica que:

> Compartilhando dores e esperanças no livre espaço público da internet, conectando-se entre si e concebendo projetos a partir de múltiplas fontes do ser, indivíduos formaram redes, a despeito de suas opiniões pessoais ou filiações organizacionais. Uniram-se.
> (...)
> Da segurança do ciberespaço, pessoas de todas as idades e condições passaram a ocupar o espaço público, num encontro às cegas entre si e com o destino que desejavam forjar, ao reivindicar seu direito de fazer história – sua história –, numa manifestação da autoconsciência que sempre caracterizou os grandes movimentos sociais[23].

O autor observa, ainda, que as redes sociais são plataformas de comunicação digital em massa, porque processam mensagens de muitos para muitos, com o potencial de alcançar milhões de receptores, assim como de autocomunicação,

PUC-Rio - Certificação Digital Nº 2112697/CA

[20] CASTELLS, Manuel. *Redes de indignação e esperança*. *Movimentos sociais na era da internet*. Tradução: Carlos Alberto Medeiros. Rio de Janeiro: Jorge Zahar, 2013, p. 10.
[21] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 93.
[22] Destaca-se, por outro lado, que vivemos hoje a era da "pós-verdade", na qual fatos objetivos são menos influentes em formar a opinião pública do que apelos à emoção, crença pessoal, e desinformação. Esse é um problema atual enfrentado pelas plataformas de redes sociais, e que atenta contra a democracia: "Foi dentro do contexto da pós-verdade que a campanha pelo Brexit, além do crescimento de campanhas de desinformação utilizadas por políticos em países como Hungria, Russia e Turquia (MCINTYRE, 2018, p. 5-6) que as fake news se apresentaram enquanto problema às democracias contemporâneas, das quais a brasileira não está à parte". KRAUS, Mariella; PANSIERI, Flávio; PAVAN, Stefano Ávila. *Desinformação, pós-verdade e democracia*: Uma análise no contexto do Estado democrático de direito. Revista Jurídica Unicuritiba. Curitiba. V.04, n.66, p.163-196 [Received/Recebido: Maio 23, 2021; Accepted/Aceito: Julho 23, 2021].
No mesmo sentido: CASTELLS, Manuel. *Ruptura:* A crise da democracia liberal. Tradução: Joana Angélica D'Avila Melo. Rio de Janeiro: Zahar, 2018; e SUNSTEIN, Cass. *#Republic: divided democracy in the age of social media*. Princeton: Princeton University Press, 2017.
[23] CASTELLS, Manuel. *Redes de indignação e esperança*. *Movimentos sociais na era da internet*. Tradução: Carlos Alberto Medeiros. Rio de Janeiro: Jorge Zahar, 2013, p. 10.

porque a produção da mensagem é decidida de forma autônoma pelo remetente. E é justamente a autocomunicação de massa que fornece a "plataforma tecnológica para a construção da autonomia do ator social, seja ele individual ou coletivo, em relação às instituições da sociedade", e é por isso que "governos têm medo da internet"[24].

Com efeito, essa nova dinâmica do debate público fez com que a censura estatal perdesse eficácia, já que, no ambiente digital, um conteúdo pode ser repostado e viralizar em questão de segundos, tornando ordens judiciais pela remoção de um conteúdo, por exemplo, muitas vezes ineficazes[25].

Por outro lado, o crescimento das redes sociais e seu uso por bilhares de pessoas ao redor do mundo também permitiu a apropriação dessas comunidades abertas para uso abusivo, com ampla disseminação de discursos de ódio, notícias falsas, e conteúdo ilegal[26]. Como não podia deixar de ser, tornou-se necessário que as plataformas de redes sociais controlassem, de alguma maneira, o conteúdo nelas postado. Começaram, então, a impor termos e condições de uso para definir os valores e normas de cada plataforma, moderando o conteúdo postado por terceiros[27]. Assim, com o tempo, as plataformas de redes sociais, que no início sugiram como meras empresas de tecnologia, passaram a exercer influência e controle sobre o discurso, tornando-se, para alguns autores, verdadeiras governantes de espaços digitais[28].

A então descentralização sobre o discurso esperada com o início da internet acabou por ser novamente centralizada, dessa vez nas mãos de poucas empresas de tecnologia. É preciso, contudo, pontuar que o controle sobre o discurso praticado por tais empresas não é comparável com aquele exercido pelos veículos de mídias

PUC-Rio - Certificação Digital Nº 2112697/CA

[24] CASTELLS, Manuel. *Redes de indignação e esperança*. Movimentos sociais na era da internet. Tradução: Carlos Alberto Medeiros. Rio de Janeiro: Jorge Zahar, 2013, p. 15.
[25] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 93.
[26] "Verifica-se que as diversas oportunidades que as redes sociais virtuais oferecem aos seus usuários, atreladas à extrema facilidade para a criação de contas pessoais, grupos e postagens, acabam contribuindo para a usurpação e a exposição injustificada de direitos de terceiros. Perfis falsos, descrições difamatórias e a exibição não consensual de imagens e informações íntimas são exemplos de utilização desses canais de comunicação que geram graves danos à pessoa humana". MORAES, Maria Celina Bodin de; TEFFÉ, Chiara Spadaccini de. *Redes sociais virtuais: privacidade e responsabilidade civil.* Análise a partir do Marco Civil da Internet. Pensar, Fortaleza, v. 22, n. 1, p. 108-146, jan./abr. 2017.
[27] BARROSO, Luna van Brussel. op. cit., p. 94.
[28] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598.

tradicionais. Com efeito, a enorme quantidade de conteúdo postado a cada instante nas plataformas não possibilita uma fiscalização e controle editorial como ocorre com os veículos de mídias tradicionais. Embora as plataformas não sejam neutras e apliquem seus termos de uso e diretrizes, a regra geral ainda é de publicação livre por qualquer usuário, salvo quando violar as regras das plataformas[29]. Assim, "há intervenção editorial, mas sob uma lógica mais excepcional, que não determina de partida qual conteúdo 'entra' na plataforma"[30].

A Era Digital permitiu, assim, a ascensão de atores privados com inédito potencial de influenciar diretamente o exercício do direito fundamental à liberdade de expressão dos usuários. Com efeito, as plataformas digitais têm poder de alcançar um enorme público, com certo controle sobre o que é publicado e disseminado em seus espaços, o que faz com que tais empresas privadas possam, unilateralmente, moldar comportamentos, crenças, e até mesmo resultados eleitorais[3132]. Cass Sunstein afirma que as redes sociais têm ainda o poder da personalização, ao criar, por meio de algoritmos, *feeds* personalizados para cada usuário, com base suas preferências, em uma espécie de curadoria do conteúdo disponibilizado para cada usuário[33].

Essa nova posição de controle das plataformas de redes sociais sobre o discurso público trouxe novos contornos ao debate da liberdade de expressão. Se antes havia um controle dual, exercido pelos Estados e pelos veículos de mídias tradicionais, o controle passou a ser triangular, como explica Jack M. Balkin. Governos e Estados seguem em uma ponta do controle, enquanto em outra permanecem os oradores. Estes, contudo, não se limitam mais aos veículos de

PUC-Rio - Certificação Digital Nº 2112697/CA

[29] BALKIN, Jack M. *Free Speech is a Triangle*. Columbia Law Review, v. 118, n. 07, p. 2024, 2018. Disponível em: <https://columbialawreview.org/content/free-speech-is-a-triangle/>. Acesso em 12 out. 2022.

[30] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 127.

[31] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 98.

[32] Disponível em: <https://www.bbc.com/portuguese/geral-37961917>. Acesso em: 12 nov. 2022.

[33] Nesse sentido, Cass Sunstein afirma que: "Com o surgimento da inteligência artificial, os algoritmos devem melhorar imensamente. Eles aprenderão muito sobre você e saberão o que você quer ou gosta, antes de você e melhor do que você. Eles até conhecerão suas emoções, novamente antes de você (...)".

Tradução livre de: "With the rise of artificial intelligence, algorithms are bound to improve immeasurably. They will learn a great deal about you, and they will know what you want or will like, before you do, and better than you do. They will even know your emotions, again before you do (...)". SUNSTEIN, Cass. #*Republic*, Divided Democracy in the Age of Social Media. Princeton: Princeton University Press, 2017.

mídias tradicionais. O advento da internet, como visto, possibilitou que cidadãos, políticos, organizações da sociedade civil e qualquer pessoa, mesmo as más intencionadas, participem de forma livre do debate[34]. Há hoje, portanto, uma terceira ponta, na qual se encontram as plataformas de redes sociais, provedores intermediários que fornecem a infraestrutura que permite publicações de conteúdo por usuários[35].

Ana Frazão e Ana Rafaela Medeiros esclarecem que, nesse contexto, o centro da proteção dos direitos individuais se deslocou da esfera pública para esfera privada, "na medida em que as plataformas digitais passam a mediar conflitos de direitos fundamentais, exercendo o poder de definição de garantias individuais relacionadas à privacidade, à honra, à liberdade de expressão etc"[36].

Essa nova dinâmica triangular trouxe novos desafios para liberdade de expressão, já que as plataformas passaram a exercer uma espécie de controle sobre o conteúdo postado por usuários com base em suas regras internas[37]. Embora tal controle não seja semelhante ao que ocorreria com os veículos de comunicação tradicionais, não se pode considerar que as plataformas oferecem um serviço neutro, sem qualquer tipo de controle[38].

Muito pelo contrário. Grande parte dos serviços oferecidos por plataformas de redes sociais está justamente na moderação do conteúdo nelas postado por terceiros. Com efeito, a internet não criou apenas um novo espaço de manifestação, mas gerou também um novo espaço de decisão quanto ao que as pessoas podem ou não dizer[39]. Afinal, até o advento do Marco Civil da Internet, que definiu o modelo de responsabilidade civil dos provedores de aplicação por conteúdo gerado por terceiros, como se verá no Capítulo 1.3.2, as empresas poderiam ser

PUC-Rio - Certificação Digital Nº 2112697/CA

[34] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 95.
[35] BALKIN, Jack M. *Free Speech is a Triangle*. Columbia Law Review, v. 118, n. 07, p. 2024, 2018. Disponível em: <https://columbialawreview.org/content/free-speech-is-a-triangle/>. Acesso em 12 out. 2022.
[36] FRAZÃO, Ana; MEDEIROS, Ana Rafaela. Responsabilidade civil dos provedores de internet: A liberdade de expressão e o art. 19 do Marco Civil. *In*: JÚNIOR; Marcos Ehrhardt; LOBO, Fabíola Albuquerque; e ANDRADE, Gustavo. *Liberdade de expressão e relações privadas*. Editora Fórum, *e-book*, p. 419.
[37] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Revie*w* 131 (2018): 1598/1670, 2018.
[38] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 119.
[39] RAMOS, Carlos Eduardo Vieira. *O direito das plataformas Digitais* - Regulação Privada da Liberdade de Expressão na Internet - Procedimento, Legitimidade e Constitucionalização. Curitiba: Juruá, 2021, p. 24.

responsabilizadas por determinado conteúdo se, após notificação extrajudicial, não o removesse. Isso gerou um controle mais rigoroso, pelas plataformas, do conteúdo postado pelos usuários, e na criação de regras compatíveis com a liberdade de expressão.

O Facebook, por exemplo, desenvolveu, em 2017, uma ferramenta que impede, automaticamente, o compartilhamento de imagens contendo pornografia não autorizada, denominada "pornografia de vingança", após condenações da empresa pela veiculação desse tipo de conteúdo[40].

A moderação de conteúdo pode ser compreendida como o "conjunto de práticas implementadas para identificar, remover e combater conteúdo ilegal ou incompatível com termos e condições fixados pelas próprias plataformas"[41]. Essa prática abrange tanto a remoção de um conteúdo, quanto a redução ou amplificação de seu acesso, ou a inclusão de informações ou redirecionamento a outro conteúdo.

Conforme o estudo "Transparência na moderação de conteúdo: tendências de governança nacional", do Instituto Referência em Internet e Sociedade[42], a gestão do conteúdo publicado nas plataformas pode ser realizada de diversas formas. Os procedimentos mais comuns são (i) a remoção completa de publicações consideradas ilegais ou inadequadas; (ii) a indisponibilização, quando o conteúdo não estiver acessível por um determinado tempo ou local; (iii) restrições ao acesso ao conteúdo, como restrições de idade; (iv) a inclusão de símbolos (*tags*), alertando sobre a natureza do conteúdo, seja patrocinado ou não, por exemplo; e (v) a classificação, utilizando algoritmos para classificar outros conteúdos como “mais apropriados” para aquela rede.

A forma pela qual as plataformas de redes sociais moderam conteúdo é geralmente estabelecida em dois principais documentos, os "termos de uso", e as "diretrizes da comunidade". O primeiro documento é geralmente mais jurídico, se apresentando como um contrato com as regras pelas quais os usuários e a plataforma interagem, as obrigações que os usuários devem aceitar como condição

PUC-Rio - Certificação Digital Nº 2112697/CA

[40] Nesse sentido: TJSP. Apelação Cível nº 1004264-50.2014.8.26.0132, Rel. Francisco Loureiro, 1ª Câmara Reservada de Direito Privado, j. em 08/07/2016.

[41] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 96.

[42] Disponível em: <https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf>. Acesso em: 19 out. 2022.

para sua participação, e as formas para resolver as disputas criadas[43]. As diretrizes da comunidade, por sua vez, é o documento que os usuários têm mais probabilidade de ler se têm alguma dúvida sobre o uso apropriado da plataforma ou algum problema com um conteúdo ou usuários ofensivos. Este documento apresenta as expectativas da plataforma sobre o que é apropriado, divulga seus princípios e lista as proibições, com graus variados de explicações para isso[44].

A atividade de moderação de conteúdo praticada pelas plataformas digitais é, portanto, necessária para assegurar a liberdade de expressão. As plataformas devem moderar conteúdo, seja para proteger um usuário de outro, um grupo de seu antagonista, seja para remover conteúdo ofensivo ou ilegal, ou seja para apresentar sua melhor forma para novos usuários e para o público em geral.

A moderação de conteúdo é necessária, ainda, para, como ensina Howard Rheingold[45], permitir o envolvimento de uma comunidade. Isso porque, com uma moderação de conteúdo eficaz, os membros engajados de uma comunidade poderão postar conteúdo relevante, que não será prejudicado por outros usuários que postem conteúdo impreciso ou abusivo, alterando o foco das comunidades em um espaço *online* seguro e compartilhado.

Nesse sentido, durante a pandemia do COVID-19, por exemplo, plataformas como Facebook, Twitter e Instagram adotaram medidas para evitar a disseminação de informações falsas. O Twitter fez alterações em suas regras de moderação de conteúdo para conseguir remover mensagens falsas sobre o coronavírus, incluindo novos critérios para remoção automática de postagens, como aquelas negando fatos científicos estabelecidos. O Facebook, por sua vez, investiu mais de 1 milhão de dólares para aprimorar sua capacidade de checagem de informações durante a pandemia, e implementou algoritmos para facilitar a busca de notícias falsas ou sensacionalistas sobre o vírus[46].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[43] GILLESPIE, Tarleton. *Custodians of the Internet:* Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, CT: Yale University Press, 2018, p. 47. Disponível em: <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em: 12 nov. 2022.

[44] Ibid.

[45] RHEINGOLD, Howard. *The Virtual Community*. Disponível em: <https://www.rheingold.com/vc/book/10.html>. Acesso em: 12 nov. 2022.

[46] Disponível em: <https://www.consumidormoderno.com.br/2020/04/01/redes-sociais-combatendo-desinformacao-coronavirus/>. Acesso em: 12 nov. 2022.

Além disso, para evitar a propagação de *fake news*[47] durante o período eleitoral das eleições de 2022 no Brasil, Facebook e Instagram passaram a marcar as postagens relacionadas às eleições, e a exibir um *link* para o Portal da Justiça Eleitoral, com o fim de ajudar as pessoas a terem acesso à informações confiáveis[48].

Por isso, Terleton Gillespie considera que a moderação de conteúdo é a principal mercadoria que as plataformas de redes sociais têm para oferecer aos usuários, se pretendem oferecer um espaço digital seguro. Com efeito, todas as plataformas moderam conteúdo, sendo essa uma atividade constante e essencial:

> Digo isso literalmente: a moderação é a essência das plataformas, é a mercadoria que elas oferecem. A essa altura do livro, isso deve estar claro. Em primeiro lugar, a moderação é uma parte surpreendentemente grande do que eles fazem, no sentido prático do dia a dia e em termos de tempo, recursos e número de funcionários que dedicam a isso. Além disso, a moderação molda como as plataformas concebem seus usuários – e não apenas aqueles que quebram regras ou buscam ajuda. Desviando parte do trabalho de moderação, por meio da sinalização, as plataformas substituem os usuários como editores amadores e policiais. A partir desse momento, os gerentes de plataforma devem, em parte, pensar, abordar e gerenciar os usuários como tal[49].

Por isso mesmo, pode-se afirmar que as plataformas de redes sociais, utilizadas por bilhares de usuários espalhados pelo mundo e caracterizadas como gigantes bancos de dados de informações desses usuários[50], se encontram

PUC-Rio - Certificação Digital Nº 2112697/CA

[47] Analisando o problema da desinformação, Alexandre Veronese e Gabriel Fonseca observam que "o dilema atual não se refere aos meios técnicos de transmissão, mas, sim, à integralidade e à confiabilidade dos conteúdos, necessitando, portanto, que formuladores de políticas públicas devam ir além. O dilema da desinformação, em verdade, se refere às garantias dos cidadãos de poder receber fluxos de informações confiáveis para poder agir socialmente, de forma racional". FONSECA, Gabriel; e VERONESE, Alexandre. Desinformação, fake news e mercado único digital: a potencial convergência das políticas públicas da União Europeia com os Estados Unidos para melhoria dos conteúdos comunicacionais. *In: Fake News e as eleições 2018*. Cadernos Adenauer, 2018, p. 42.

[48] Disponível em: <https://www.tse.jus.br/comunicacao/noticias/2021/Dezembro/contra-fake-news-instagram-e-facebook-colocam-avisos-em-postagens-sobre-eleicoes-2022>. Acesso em: 12 nov. 2022.

[49] Tradução livre de: "I mean this literally: moderation is the essence of platforms, it is the commodity they offer. By this point in the book, this should be plain. First, moderation is a surprisingly large part of what they do, in a practical, day- to- day sense, and in terms of the time, resources, and number of employees they devote to it. Moreover, moderation shapes how platforms conceive of their users—and not just the ones who break rules or seek help. By shifting some of the labor of moderation, through flagging, platforms deputize users as amateur editors and police. From that moment, platform managers must in part think of, address, and manage users as such". GILLESPIE, Tarleton. *Custodians of the Internet*: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (New Haven, CT: Yale University Press, 2018), 5, p. 14. Disponível em: <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em 12 out. 2022.

[50] Como esclarece Shoshana Zuboff ao tratar do capitalismo de vigilância, as grandes empresas de tecnologia alcançaram o poder que têm hoje através de um processo de extração de dados de usuários, que constituem sua principal mercadoria, e tornam possível influenciar o comportamento de usuários. ZUBOFF, Shoshana. *A era do capitalismo de vigilância*: a luta por um futuro humano na nova fronteira do poder; tradução George Schlesinger. 1ª ed. Rio de Janeiro: Intrínseca, 2020.

atualmente como importantes agentes na estruturação do discurso público, atuando como definidoras de normas e julgadoras de disputas, em âmbito administrativo[51]. Por isso mesmo, as plataformas privadas e as regras por elas criadas têm papel fundamental sobre o discurso público[52]. Especificamente em relação ao Facebook, esclarece Kate Klonick:

> A fim de implementar as suas Regras de Comunidade, o Facebook desenvolveu um sistema burocrático imenso para moderar conteúdo de usuários e adjudicar disputas decorrentes desse conteúdo. Devido ao enorme volume de conteúdo publicado diariamente, o Facebook não consegue fazer e não faz monitoramento proativo de violações a suas regras. Detecções automatizadas de violações são sofisticadas e bem-sucedidas para diversos tipos de conteúdo visual (como pornografia infantil), mas menos eficazes para conteúdo escrito que apresenta 'desafios linguísticos diferenciados' (como assédio e discurso de ódio). Como consequência, a plataforma ainda depende de usuários que reativamente denunciem discurso que pode violar as suas regras. Conteúdo reportado por usuários é colocado em uma fila para revisão humana efetivada por moderadores de conteúdo[53].

Embora, de certo modo, os termos e condições de uso muitas vezes apenas reflitam aquilo já definido em lei (como assédio e pornografia infantil), eles ainda podem criar requisitos adicionais que são formulados, muitas vezes, de forma genérica, dificultando que se saiba antecipadamente como e porque serão aplicados em casos concretos[54].

PUC-Rio - Certificação Digital Nº 2112697/CA

Ademais, como observa Cass Sunstein, é por meio dos dados coletados dos usuários que as plataformas interferem na construção e na difusão da informação, personalizando a informação que será visualizada por cada usuário. SUNSTEIN, Cass. *#Republic*, Divided Democracy in the Age of Social Media, Princeton: Princeton University Press, 2017.
No mesmo sentido: FRAZÃO, Ana. *Plataformas Digitais e o Negócio de Dados*: Necessário Diálogo entre o Direito da Concorrência e a Regulação dos Dados. RDP, Brasília, Volume 17, n. 93, 58-81, maio/jun. 2020.

[51] GILLESPIE, Tarleton. *Custodians of the Internet:* Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (New Haven, CT: Yale University Press, 2018), 5, p. 14. Disponível em: <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em 12 out. 2022.

[52] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 96.

[53] Tradução livre de: "In order to implement the Community Standards, Facebook has developed an immense bureaucratic system to moderate user content and adjudicate disputes arising from that content. Because an enormous volume of content is posted every day, Facebook cannot and does not proactively police all violations of its rules. Automated detection of violations is quite sophisticated and successful for various types of visual content (such as child pornography) but less so for written content that poses “nuanced linguistic challenges” (such as harassment and hate speech).123 As a result, the platform still relies on users to reactively flag speech that might violate its rules". KADRI, Thomas E., KLONICK, Kate. *Facebook v. Sullivan: public figures and newsworthiness in online speech*. Southern California Law Review, v. 93, p. 37-99, 2019, p. 59.

[54] HUMAN RIGHTS COMMITTEE. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. 11 mai. 2016. UN Doc A/HRC/32/38. Disponível em: <https://undocs.org.en/A/HRC/32/38>. Acesso em: 12 nov. 2022.

De fato, há falta de transparência e devido processo legal não só na forma como tais regras são aplicadas, mas na própria existência dessas regras[55]. Muitas vezes um usuário tem um conteúdo removido por uma plataforma sem saber as razões para isso, e sem que tenha a chance de questionar tal decisão. Essa falta de transparência pode fazer com que a fiscalização de conteúdos seja feita de forma inconsistente, podendo lesar minorias e levar à remoção de conteúdo muitas vezes lícito, prejudicando a liberdade de expressão[56].

Não se pode deixar de considerar, contudo, que a atividade de moderação de conteúdo na internet não é uma tarefa simples. Muito pelo contrário. Moderar conteúdo é desafiador não apenas porque demanda recursos, mas principalmente porque exige que as plataformas tomem decisões difíceis sobre aspectos que muitas vezes não são facilmente identificáveis como ilícitos ou ilegais. Ora, qual é a diferença entre sexualmente explícito e pornográfico? Quando uma imagem do corpo humano é artística, educacional ou lasciva? As representações da violência ficcional são meramente divertidas ou psicologicamente prejudiciais[57]? Essas questões estão sempre presentes nos esforços para moderar conteúdo questionável e dependem não apenas de diferentes valores e ideologias, mas também de teorias contestadas de impacto psicológico e políticas culturais diferentes.

Assim, embora a moderação de conteúdo seja necessária no âmbito digital, sob pena de se permitir a proliferação de conteúdo indesejável e ambientes tóxicos, a forma opaca pela qual ela acabou sendo desenvolvida pelas plataformas apresenta problemas que merecem ser enfrentados. Afinal, a maior parte do controle sobre o discurso público atual é feito por poucas plataformas, que exercem, de forma simultânea, funções legislativas, executivas e jurídicas[58]. São elas que definem as regras sobre o conteúdo permitido, que aplicam essas regras, e que também,

PUC-Rio - Certificação Digital Nº 2112697/CA

[55] Com efeito, a despeito, por exemplo, da obrigação da Lei Geral de Proteção de Dados (Lei Federal nº 13.709, de 2018) de que plataformas adotem o *privacy by design* (framework que tem como proposta central incorporar a privacidade e a proteção de dados pessoais em todos os projetos desenvolvidos por uma organização, desde a sua concepção), muitas vezes essas regras são construídas em termos genéricos, incompreensíveis aos usuários, que, por não conseguir compreender a regra, também não conseguirá disputar uma decisão tomada com base nela.

[56] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 97.

[57] GILLESPIE, Tarleton. *Custodians of the Internet*: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (New Haven, CT: Yale University Press, 2018), 5, p. 14. Disponível em: <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em 12 out. 2022.

[58] BARROSO, Luna van Brussel. op. cit., p. 98.

internamente, revisam recursos[59]. Além disso, a falta de transparência sobre aplicação e controle de tais regras dificulta um controle público e uma *accountability* sobre a atuação das plataformas.

Como sumariza Luna Barroso, os principais pontos de preocupação em relação à moderação de conteúdo das plataformas digitais envolvem:

> (i) regras vagas, que não indicam claramente o que caracteriza violação; (ii) possibilidade de aplicação inconsistente e não isonômica dos termos e condições, com impacto prejudicial sobre minorias; (iii) falta de análise do contexto em que as manifestações foram proferidas, que são fundamentais para a definição da licitude ou ilicitude do discurso; (iv) dificuldade de adoção de medidas proporcionais de combate a conteúdo ilícito ou danoso, incluindo a disseminação de desinformação e propaganda que ameaçam a confiança pública nos veículos de mídia e nas instituições governamentais; (v) uso de algoritmos e mecanismos automatizados de remoção ou filtragem de conteúdo; (vi) possibilidade de que os mecanismos de denúncia de conteúdo sejam subvertidos por usuários e usados para silenciar conteúdo lícito; (vii) ausência de notificação ao usuário afetado sobre a remoção de conteúdo e do fundamento para tanto; e (viii) insuficiência dos mecanismos de recurso interno às plataformas[60].

Tais dificuldades levaram à necessidade de uma regulação da moderação de conteúdo. Afinal, se nos veículos tradicionais era possível saber o que estava sendo veiculado aos assinantes e as escolhas editoriais de cada veículo eram sujeitas ao escrutínio público e até mesmo à responsabilidade civil, como em caso de calúnia e difamação no Brasil, isso não vale para as plataformas digitais[61]. Nesse novo universo o controle é mais sutil, opaco, e menos transparente, pois, como se verá no Capítulo 1.4, é introduzido muitas vezes por algoritmos de filtros de difícil detecção[62].

Ademais, os novos agentes na estruturação do discurso público possuem mais informações pessoais, incluindo preferências políticas, religiosas e sexuais dos usuários[63], o que acaba potencializando o poder de influenciar comportamentos ou mesmo vender dados pessoais dos usuários para terceiros que efetuam tal influência[64]. Esse controle desenvolvido pelas plataformas pode ser limitado por

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[59] KADRI, Thomas E., KLONICK, Kate. *Facebook v. Sullivan: public figures and newsworthiness in online speech*. Southern California Law Review, v. 93, p. 37-99, 2019, p. 94.
[60] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 98.
[61] FUKUYAMA, Francis, et al. *Report of the working group on platform scale*. Stanford Program on Democracy and the Internet, 2020. Disponível em: <https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale>. Acesso em: 12 nov. 2022.
[62] BARROSO, Luna van Brussel. op. cit, p. 139.
[63] FUKUYAMA, Francis, et al. op. cit.
[64] BARROSO, Luna van Brussel. op. cit.

meio de regulação, que deve considerar essa assimetria de informações com maiores deveres de transparência sobre as plataformas.

1.2. A liberdade de expressão no ordenamento jurídico brasileiro e o entendimento do Supremo Tribunal Federal sobre esse direito fundamental

No Brasil, a liberdade de expressão é assegurada pelos artigos 5º, inciso IV, e 220, da Constituição Federal, que elevam a liberdade de pensamento e sua expressão ao nível de garantia fundamental. A liberdade de expressão é um "relevante aspecto da autonomia do indivíduo, concebida, numa perspectiva kantiana, como o centro da dignidade da pessoa humana (...) não pode haver dúvida de que a liberdade de expressão é crucial para a participação do cidadão no processo democrático"[65]. Como afirmou a Ministra Carmen Lúcia "a liberdade maior que se tem é a da expressão"[66].

Aline Osório observa que a liberdade de expressão tem posição de destaque na Constituição de 1988, principalmente em resposta "ao trauma do período autoritário, tendo sido gestadas como parte do processo de redemocratização do Brasil, de modo a garantir a não repetição da censura característica da ditadura militar"[67]. A proteção especial conferida à liberdade de expressão se justifica por 3 fatores: (i) a busca da verdade; (ii) a garantia da dignidade humana e da autonomia individual[68]; e (iii) a realização da democracia[69].

O Supremo Tribunal Federal ("STF") reconhece que a liberdade de expressão é tão importante que deve ser considerada uma "garantia preferencial" quando confrontada com direitos fundamentais concorrentes, "em razão da estreita relação com outros princípios e valores fundantes, como a democracia, a dignidade da pessoa humana, a igualdade"[70]. "[Q]uando se tem um conflito possível entre a

[65] BRASIL. STF. Voto do Min. Luiz Fux na ADPF nº 187/DF, Rel. Min. Celso de Mello, j. em 15/6/2011.
[66] BRASIL. STF. Voto da Min. Carmen Lúcia na ADPF nº 187/DF, Rel. Min. Celso de Mello, j. em 15/6/2011.
[67] OSORIO, Aline. *Direito eleitoral e liberdade de expressão*. Belo Horizonte: Fórum, 2017, p. 39.
[68] SARMENTO, Daniel. A Liberdade de Expressão e o Problema do "Hate Speech". *In: Livres e iguais:* estudos de Direito Constitucional. Rio de Janeiro: Lumen Juris, 2006, p. 242.
[69] OSORIO, Aline. op. cit., p. 67.
[70] BRASIL. STF. RE nº 685.493, Rel. Min. Marco Aurélio, j. em 20/11/2014.

PUC-Rio - Certificação Digital Nº 2112697/CA

liberdade e sua restrição, deve-se defender a liberdade. O preço do silêncio para a saúde institucional dos povos é muito mais alto do que o preço da livre circulação das ideias"[71]. Assim, não é surpresa que o direito à liberdade de expressão seja um direito humano fundamental reconhecido internacionalmente[72].

Em 2019, ao analisar o pedido do Ministério Público Federal ("MPF") de suspensão liminar ("SL") nº 1.248 contra a decisão da presidência do Tribunal de Justiça do Rio de Janeiro ("TJRJ") que havia declarado a legalidade da apreensão, por fiscais da prefeito do Rio de Janeiro, de revistas na Bienal do Livro por conterem a imagem de um beijo gay na capa, diante da suposta necessidade de proteção de crianças e adolescentes, o Ministro Dias Toffoli deferiu o pedido liminar para suspender a decisão da presidência do TJRJ. Na oportunidade, o Ministro reconheceu que o "Supremo Tribunal Federal tem construído uma jurisprudência consistente em defesa da liberdade de expressão", afirmando que "a democracia somente se firma e progride em um ambiente em que diferentes convicções e visões de mundo possam ser expostas, defendidas e confrontadas umas com as outras, em um debate rico, plural, e resolutivo"[73]. Reconheceu, portanto, o caráter instrumental da liberdade de expressão para democracia, bem como seu status de direito humano universal. O Ministro afirmou, ainda, que a

PUC-Rio - Certificação Digital Nº 2112697/CA

[71] BRASIL. STF. Voto do Min. Menezes Direito na ADPF nº 130, Rel. Min. Carlos Ayres Britto, j. em 30/4/2009.

[72] Veja-se, por exemplo, a Declaração Universal dos Direitos Humanos: "Todo ser humano tem direito à liberdade de opinião e expressão; esse direito inclui a liberdade de, sem interferência, ter opiniões e de procurar, receber e transmitir informações e ideias por quaisquer meios e independentemente de fronteiras." (artigo 19); *Pacto Internacional sobre Direitos Civis e Políticos*: "Toda pessoa terá direito à liberdade de expressão; esse direito incluirá a liberdade de procurar, receber e difundir informações e ideias de qualquer natureza, independentemente de considerações de fronteiras, verbalmente ou por escrito, em forma impressa ou artística, ou por qualquer outro meio de sua escolha.," e isso só pode ser restringido por lei expressa, que visa a "que se façam necessárias, em uma sociedade democrática, no interesse da segurança nacional, da segurança ou da ordem pública, ou para proteger a saúde ou a moral pública ou os direitos e as liberdades das demais pessoas."; *Convenção Americana sobre Direitos Humanos:* o direito à liberdade de pensamento e expressão incluem o direito de "buscar, receber e difundir informações e ideias de toda natureza, sem consideração de fronteiras, verbalmente ou por escrito, ou em forma impressa ou artística, ou por qualquer outro processo de sua escolha." (artigo 13); *Carta dos Direitos Fundamentais da União Europeia*: o Direito à liberdade de expressão e informação deve compreender "a liberdade de opinião e a liberdade de receber e de transmitir informações ou ideias, sem que possa haver ingerência de quaisquer poderes públicos e sem consideração de fronteiras" (artigo 11); *Carta Africana dos Direitos Humanos e dos Povos:* "E Toda pessoa tem direito à informação." e "Toda pessoa tem direito de exprimir e de difundir as suas opiniões no quadro das leis e dos regulamentos." (artigo 9°).

[73] BRASIL. STF, SL nº 1.248, Rel. Min. Dias Toffoli, DJe 10/09/2019.

"liberdade de expressão está amplamente protegida em nossa ordem constitucional", "sendo um dos grandes legados da Carta Cidadã"[74].

Em outro precedente, de 2020, o STF julgou a Reclamação nº 38.782, proposta pela Netflix para suspender os efeitos da decisão de desembargadora do TJRJ que havia determinado a suspensão da exibição de obra artística que supostamente violaria valores cristões (o "Especial de Natal Porta dos Fundos: a primeira tentação de Cristo"). O Ministro Dias Tofolli concedeu a medida liminar, suspendendo os efeitos da decisão do TJRJ. A reclamação foi então distribuída ao Ministro Gilmar Mendes, e, no julgamento do mérito, a Segunda Turma confirmou a liminar, julgando procedente o pedido da Netflix. Na oportunidade, o Ministro Gilmar Mendes afirmou que a obra "não incita violência contra grupos religiosos, mas constitui mera crítica, realizada por meio de sátira, a elementos caros ao Cristianismo"[75]. Destacou, ainda, que "a proibição de divulgação de determinado conteúdo deve se dar apenas em casos excepcionalíssimos, como na hipótese de configurar ocorrência de prática ilícita, de incitação à violência ou à discriminação, bem como propagação de discurso de ódio"[76].

Em 2021, o STF julgou o caso Aida Curi (RE nº 1.010.606), que envolvia o pedido de reconhecimento de um direito ao esquecimento. No julgamento, o Supremo, dando novamente destaque especial à liberdade de expressão, firmou a tese de que tal direito é incompatível com a Constituição Federal[77], mas que "eventuais excessos ou abusos no exercício da liberdade de expressão e de informação devem ser analisados caso a caso, a partir dos parâmetros constitucionais"[78]. O voto do Ministro Relator Dias Toffoli observou que "tanto quanto possível, portanto, deve-se priorizar: o complemento da informação, em vez de sua exclusão; a retificação de um dado, em vez de sua ocultação; o direito de resposta, em lugar da proibição ao posicionamento"[79].

PUC-Rio - Certificação Digital Nº 2112697/CA

[74] Ibid.
[75] BRASIL. STF, Rcl. nº 38.782, Rel. Min. Gilmar Mendes, DJe 23/02/2021.
[76] Ibid.
[77] "É incompatível com a Constituição a ideia de um direito ao esquecimento, assim entendido como o poder de obstar, em razão da passagem do tempo, a divulgação de fatos ou dados verídicos e licitamente obtidos e publicados em meios de comunicação social analógicos ou digitais".
[78] BRASIL. STJ, RE nº 1.010.606, Rel. Min Dias Toffoli, DJe 19/05/2021.
[79] Ibid.

Ressalta-se, contudo, que embora tenha posição preferencial, a liberdade de expressão não é um direito absoluto[80] "e não pode ser tida como um fetiche ou como um valor que deva sempre prevalecer sobre os demais"[81]. A liberdade de expressão, como todo direito fundamental, pode ser limitada em casos de conflito com outros valores e direitos constitucionalmente tutelados[82].

O direito à liberdade de expressão está intimamente ligado ao direito fundamental da liberdade de comunicação, previsto no inciso IX do artigo 5º, e no artigo 220 da Constituição Federal. Este protege os direitos dos cidadãos de se comunicarem livremente utilizando a forma, o meio ou o método que escolherem, independentemente de censura ou permissão. A possibilidade de exercer plenamente a liberdade de comunicação não se dá apenas pela proteção do ato de se comunicar – o qual os cidadãos podem exercer escolhendo seus meios de comunicação sem qualquer interferência estatal –, mas também pela proteção adequada da infraestrutura que permite com que tal comunicação ocorra. Conforme explicado pelo Ministro Celso de Mello, a liberdade de comunicação significa que "(...) o Estado não pode dispor de poder algum sobre a palavra, sobre as ideias e sobre os modos de manifestação"[83].

PUC-Rio - Certificação Digital Nº 2112697/CA

1.3 Regime legal que possibilitou o exercício da atividade de moderação de conteúdo

O presente capítulo apresenta, inicialmente, o regime legal da Seção 230, do *Communications Decency Act* ("CDA"), "Ato da Decência das Comunicações", dos Estados Unidos. Tal Seção conferiu duas imunidades às plataformas de redes sociais: (i) imunidade sobre o conteúdo postado por terceiros; e (ii) imunidade sobre a moderação de conteúdo feita pelas redes sociais. Dessa forma, as plataformas não

[80] "A liberdade de expressão, portanto, não está imune a restrições nem assume posição hierárquica de superioridade quando em conflito com outros direitos fundamentais, exigindo-se, ao contrário, um cuidadoso balanceamento dos bens jurídicos contrapostos". FRAZÃO, Ana; MEDEIROS, Ana Rafaela. Responsabilidade civil dos provedores de internet: A liberdade de expressão e o art. 19 do Marco Civil. *In*: JÚNIOR; Marcos Ehrhardt; LOBO, Fabíola Albuquerque; e ANDRADE, Gustavo. *Liberdade de expressão e relações privadas*. Editora Fórum, *e-book*, p. 423.
[81] OSORIO, Aline. *Direito eleitoral e liberdade de expressão*. Belo Horizonte: Fórum, 2017, p. 116.
[82] BARROSO, Luís Roberto. *Colisão entre liberdade de expressão e Direitos da personalidade. Critérios de ponderação. Interpretação Constitucionalmente adequada do Código Civil e da Lei de Imprensa*. Revista de Direito Administrativo, v. 235, p. 1-36, 2001. Disponível em: <https://bibliotecadigital.fgv.br/ojs/index.php/rda/article/view/45123>. Acesso em: 25 nov. 2022.
[83] BRASIL. STF. ADPF nº. 187/DF, Tribunal Pleno, Rel. Min. Celso de Mello, j. em 15/06/2011.

podem ser responsabilizadas pelo conteúdo de terceiros, e nem por moderar esse conteúdo, caso viole suas políticas internas. Considera-se que foi a Seção 230, do CDA, que permitiu que as plataformas alcançassem a posição que têm hoje como controladoras do discurso digital, diante da ampla imunidade conferida às empresas.

Será visto, ainda, como o legislador brasileiro se posicionou sobre o tema, por meio do regime de responsabilidade civil adotado pelo Marco Civil da Internet. De forma semelhante à Seção 230, o Marco Civil da Internet também isenta provedores de aplicação de responsabilidade civil pelo conteúdo gerado por terceiro, salvo em caso de descumprimento de ordem judicial específica. Embora não seja expresso em relação à moderação de conteúdo, entende-se que o Marco Civil da Internet não proíbe essa atividade, de modo que as plataformas também não podem ser responsabilizadas por moderar conteúdo com base em suas regras internas, desde que observem determinados deveres procedimentais, como transparência e possibilidade de resposta.

PUC-Rio - Certificação Digital Nº 2112697/CA

1.3.1 Estados Unidos: Seção 230 do *Communications Decency Act* ("CDA")

Em 1996, foi aprovado nos Estados Unidos o *Communications Decency Act* ("CDA"), "Ato da Decência das Comunicações". Esse nome surgiu diante da preocupação com o fato de que crianças e adolescentes estavam tendo acesso a conteúdo impróprio para sua idade, entendendo-se necessária a criação de certos filtros e barreiras para evitar isso.

O legislador americano entendeu que a internet propiciava não somente a promessa de uma comunicação global e instantânea, em que se poderia mandar uma mensagem para alguém do outro lado do globo que no segundo seguinte a receberia, mas também um ambiente propício para que abusos fossem cometidos.

Na chamada web 1.0, muitos dos sites mais populares e promissores da internet eram aqueles que se caracterizavam apenas por ser, por exemplo, um portal noticioso. Já na web 2.0 ou web colaborativa, a internet deixou de ser uma sucessão de portais noticiosos e se tornou dominada por plataformas em que se tem acesso a

conteúdo criado pelos próprios usuários[84]. Ou seja, a internet deixou de ser uma versão digital de veículos jornalísticos que existiam em meio impresso, permitindo que novas plataformas de acesso a conteúdo surgissem, com infraestruturas tecnológicas acessadas globalmente em que um usuário pode interagir, compartilhar, curtir e assistir a conteúdo que é era gerado por essas plataformas, mas sim pelos seus usuários, uma infinidade de pessoas que se juntavam nessas novas praças públicas virtuais.

O legislador americano entendeu esse momento, e antes mesmo de se consagrar a expressão "web 2.0", editou o CDA, que foi resultado de uma avaliação pragmática do Congresso americano, motivado por dois precedentes. O primeiro, Cubby, Inc. v. CompuServe, Inc., envolveu a publicação de difamação na plataforma CompuServe[85]. O Tribunal do Distrito Sul de Nova York, em 1991, entendeu que a CompuServe não poderia ser responsabilizada pelo conteúdo difamatório, pois não teria, como um intermediário, qualquer controle editorial sobre o conteúdo postado em seu site, sendo mero distribuidor, e não editor do conteúdo.

Quatro anos depois surgiu o segundo caso, Stratton Oakmont, Inc. v. Prodigy Services Co[86]. Nesse, o Tribunal Superior de Nova York, em 1995, entendeu que a empresa intermediária Prodigy era responsável, como editora, por todo conteúdo disponibilizado em seu site, porque ativamente agia para deletar determinados conteúdo, com verdadeiro controle editorial. O caso criou um grande desincentivo para que plataformas digitais expandissem suas atividades de moderação de conteúdo para frear conteúdo abusivo ou ofensivo[87].

Pouco tempo depois, o senador James Exon introduziu o CDA, que visava regular conteúdo obsceno postado *online*, tornando ilegal enviar ou mostrar conteúdo indecente para menores de idade. Grande parte dessa legislação foi declarada inconstitucional por não fazer um balanceamento correto entre a demanda

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[84] Disponível em: <https://www.geeksforgeeks.org/web-1-0-web-2-0-and-web-3-0-with-their-difference/>. Acesso em: 25 nov. 2022.

[85] CUBBY, Inc. v. CompuServe, Inc. U.S. District Court for the Southern District of New York - 776 F. Supp. 135 (S.D.N.Y. 1991) October 29, 1991.

[86] STRATTON OAKMONT, Inc. v. Prodigy Services Co. Supreme Court, Nassau County, New York, Trial IAS Part 34. May 24, 1995.

[87] KLONICK, Kate. "T*he New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598, p. 1603. Disponível em: <https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf>. Acesso em: 25 nov. 2022.

de proteção de menores e os ônus que eram colocados aos agentes da internet, inclusive com determinações que naquele momento eram tecnologicamente impossíveis. Alguns dispositivos permaneceram válidos, incluindo a Seção 230, principal dispositivo que regulamenta a responsabilidade dos intermediários na internet[88].

A Seção 230 traz duas imunidades às plataformas digitais. Em seu *caput* é dito que provedores não serão responsabilizados como se fossem editores ou "publicadores" do conteúdo de terceiros. Ou seja, prevê que não se deve responsabilizar a empresa que explora uma plataforma digital como se fosse a editora do conteúdo nela postado por terceiros, porque ela apenas disponibilizaria aquele espaço para que o conteúdo fosse postado.

A Seção 230 traz também uma segunda imunidade, chamada imunidade do "bom samaritano". Se um provedor identifica um conteúdo como sendo obsceno ou danoso pode agir para removê-lo, pois também não será responsabilizado por isso. A lógica do legislador americano foi dar imunidade para as plataformas retirando sua responsabilidade sobre conteúdo de terceiros, ainda que ajam para remove-lo.

O regime de responsabilização civil por conteúdo gerado por terceiros da Seção 230 impulsionou o desenvolvimento de provedores de aplicações e a própria expansão da internet, permitindo que se desenvolvessem plataformas como as conhecidas atualmente.

No entanto, as imunidades conferidas pela Seção 230 passaram a ser vistas como excessivas, especialmente após o surgimento das grandes redes sociais[89]. Quando, por exemplo, o Facebook e o Twitter passaram a remover conteúdo postado pelo ex-presidente Donald Trump, as plataformas foram acusadas de praticar censura[90], tendo sido articuladas tentativas infrutíferas de revogação da Seção 230. O então presidente chegou a editar uma Ordem Executiva (nº 13925, de junho de 2020), estabelecendo diretrizes para alterar a Seção 230, com vistas ao combate e à prevenção de censura nas redes sociais[91]. Segundo a referida Ordem

[88] Disponível em: <https://www.law.cornell.edu/uscode/text/47/230#fn002009>. Acesso em: 14 nov. 2022.
[89] WU, Tim. *Is the first amendment obsolete?* L. Rev. 547, p. 548. 2018. Disponível em:<https://repository.law.umich.edu/ mlr/vol117/iss3/4>. Acesso em: 28 nov. 2022.
[90] Disponível em: <https://www.opendemocracy.net/pt/censura-twitter-facebook-donald-trump-consequencia-democracia/>. Acesso em: 12 nov. 2022.
[91] VENTURI, Thaís G. Pascoaloto. *Redes Sociais: Platforms ou Publishers? - Parte I*. Disponível em: <https://www.migalhas.com.br/coluna/direito-privado-no-common-law/339965/redes-sociais-platforms-ou-publishers--parte-i>. Acesso em: 23 nov. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

Executiva, "Twitter, Facebook, Instagram e Youtube exercem um poder imenso, senão sem precedentes, de moldar a interpretação de eventos públicos; censurar, excluir ou desaparecer informações: e controlar o que as pessoas veem ou não veem"[92].

Jeff Kosseff comenta que, desde 2019, mais de 25 propostas legislativas foram apresentadas para alterar ou revogar a Seção 230, na tentativa de alterar o regime de responsabilidade e obrigações das plataformas digitais[93]. Recentemente, o presidente Joe Biden alegou que a Seção 230 deveria ser revogada diante da ampla desinformação veiculada *online*[94]. Especificamente em relação ao Facebook, afirmou que: "Não há nenhum impacto editorial no Facebook. Nenhum. Nenhum. É irresponsável. É totalmente irresponsável"[95]. Percebe-se, com isso, uma tentativa de limitar o poder dessas empresas privadas na atividade de moderação de conteúdo, que gera impactos diretos no discurso público e na liberdade de expressão[96].

Nesse sentido, Terleton Gillespie esclarece que a ampla imunidade da Seção 230 deveria vir acompanhada de obrigações. Para o autor, as plataformas digitais que exercem moderação de conteúdo deveriam observar padrões mínimos ou se comprometer com alguma forma de transparência, ou mesmo fornecer estruturas específicas para contestar decisões das plataformas. Sem isso, as plataformas continuarão a gozar de ampla imunidade, sem qualquer responsabilidade[97].

1.3.2 Brasil: Artigo 19 da Lei Federal nº 12.965/2014 ("Marco Civil da Internet")

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[92] Tradução livre de: "Twitter, Facebook, Instagram, and Youtube wield immense, if not unprecedented, power to shape the interpretation of public events; to censor, delete, or disappear information: and to control what people see or do not see". Federal Register, Vol. 85, N. 106, Tuesday, June 2, 2020. Presidential Documents. Executive Order 13925 - Preventing Online Censorship.

[93] KOSSEFF, Jeff. *A user's guide to Section 230, and a legislator's guide to amending it (or not).* Berkeley Technology Law Journal, v. 37, nº 2, 2022.

[94] Disponível em: <https://www.conjur.com.br/2022-mai-04/direito-digital-secao-230-cda-artigo-19-marco-civil-internet#_ftn6>. Acesso em: 23 nov. 2022.

[95] Tradução livre de: "There is no editorial impact at all on Facebook. None. None whatsoever. It's irresponsible. It's totally irresponsible". Disponível em: <https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html>. Acesso em: 23 nov. 2022.

[96] KOSSEFF, Jeff. Op. cit.

[97] GILLESPIE, Tarleton. *Custodians of the Internet:* Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (New Haven, CT: Yale University Press, 2018), 5, p. 52. Disponível em: <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em 12 out. 2022.

O Marco Civil da Internet é a principal lei sobre uso da internet no Brasil e tem compromisso explícito com a preservação da liberdade de expressão[98]: (i) "A disciplina do uso da internet no Brasil tem como fundamento o respeito à liberdade de expressão" (artigo 2º, *caput*); (ii) "A disciplina do uso da internet no Brasil tem os seguintes princípios: garantia da liberdade de expressão, comunicação e manifestação de pensamento, nos termos da Constituição Federal" (artigo 3º, I); e (iii) "A garantia do direito à privacidade e à liberdade de expressão nas comunicações é condição para o pleno exercício do direito de acesso à internet" (artigo 8º).

A proteção desses direitos pelo Marco Civil da Internet reflete a intenção do legislador de subordinar o arcabouço legal do uso da internet no Brasil a um sistema abrangente de princípios destinados a proteger os direitos fundamentais dos usuários.

O Marco Civil da Internet trata em especial de dois tipos de provedores: (i) os dedicados a prover o acesso à Internet, como a VIVO e outras operadoras de telefonia (de conexão), e (ii) aqueles que disponibilizam as mais diversas aplicações na rede, como Facebook e Twitter (de aplicação). Importa para o presente estudo tratar do regime de responsabilidade civil dos provedores de aplicação.

Antes do advento do Marco Civil da Internet, doutrina e jurisprudência se dividiam em três entendimentos distintos. O primeiro era pela não responsabilização do provedor de aplicação, por ser este mero intermediário entre usuário e vítima. Vale ressaltar que essa é a corrente adotada nos Estados Unidos, por meio da Seção 230, do CDA, que isenta provedores de aplicação de responsabilidade por atos de terceiros[99].

Essa regra, contudo, tem exceções mesmo no direito americano, como no caso de conteúdo que infringe direitos autorais. Para direitos autorais, os Estados Unidos adotam o mecanismo de *notice and take down* previsto no *Digital Millenium*

PUC-Rio - Certificação Digital Nº 2112697/CA

[98] "De fato, o Marco Civil realizou uma valorização da liberdade de expressão, estando tal opção legislativa de acordo com recentes posicionamentos do Supremo Tribunal Federal. Todavia, isso não significa que o intérprete deva atribuir à liberdade de expressão a condição de direito absoluto, imune a qualquer limite, nem mesmo que deva estabelecer uma espécie de hierarquia prévia entre as normas constitucionais (SARLET, 2015)". MORAES, Maria Celina Bodin de; TEFFÉ, Chiara Spadaccini de. *Redes sociais virtuais: privacidade e responsabilidade civil. Análise a partir do Marco Civil da Internet*. Pensar, Fortaleza, v. 22, n. 1, p. 108-146, jan./abr. 2017.

[99] Disponível em: <https://www.law.cornell.edu/uscode/text/47/230>. Acesso em: 8 Jan. 2022.

Copyright Act ("DMCA")[100]. Por esse mecanismo, em resumo, o provedor recebe uma notificação do suposto detentor dos direitos autorais, e com isso deve tirar o conteúdo do ar imediatamente. Após a remoção do conteúdo, seu autor também tem a oportunidade de efetuar uma contranotificação[101].

O segundo entendimento era pela responsabilidade objetiva do provedor, fundada no conceito de risco de atividade ou no defeito do serviço, considerando a relação de consumo estabelecida entre o usuário e o provedor. Tal tese encontrou suporte no artigo 927, parágrafo único, do Código Civil/02, e foi num primeiro momento o entendimento dominante do Superior Tribunal de Justiça ("STJ"), que acabava impondo ao provedor de aplicação o dever de fiscalização prévia do conteúdo.

Em seguida, o STJ passou a entender o contrário, chegando ao terceiro entendimento, pela inaplicabilidade do artigo 927, parágrafo único, do Código Civil/02, e pela consequente responsabilidade subjetiva dos provedores de aplicação por danos causados por terceiros em suas aplicações. Com efeito, o STJ passou a considerar que o dano moral decorrente de mensagens com conteúdo ofensivo inseridas por terceiros não é um risco inerente à atividade do provedor, pois não é um dever seu controlar previamente o conteúdo postado por usuários[102]. Logo, antes de o Marco Civil da Internet entrar em vigor, a jurisprudência pacífica do STJ era pela responsabilidade civil subjetiva do provedor de aplicação, por culpa

PUC-Rio - Certificação Digital Nº 2112697/CA

[100] Disponível em: <https://www.copyright.gov/legislation/dmca.pdf>. Acesso em: 8 Jan. 2022.

[101] Se isso ocorrer, o provedor de aplicação deverá substituir o material removido e deixar de desativar o acesso a ele no prazo entre 10 a 14 dias úteis após receber a contra notificação. É uma espécie de responsabilidade civil *ex post*, posterior ao início da produção do dano, voltada a impedir que o dano se propague.

[102] "No que tange à fiscalização do conteúdo das informações postadas por cada usuário, não se trata de atividade intrínseca ao serviço prestado, de modo que não se pode reputar defeituoso, nos termos do art. 14 do CDC, o site que não examina e filtra o material nele inserido.

(...) exigir dos provedores de conteúdo o monitoramento das informações que veiculam traria enorme retrocesso ao mundo virtual, a ponto de inviabilizar serviços que hoje estão amplamente difundidos no cotidiano de milhares de pessoas. A medida, portanto, teria impacto social e tecnológico extremamente negativo. (...)

O dano moral decorrente de mensagens com conteúdo ofensivo inseridas no site pelo usuário não constitui risco inerente à atividade dos provedores de conteúdo, de modo que não se lhes aplica a responsabilidade objetiva prevista no art. 927, parágrafo único, do CC/02". BRASIL. STJ, REsp, 1.186.616/MG, 3ª Turma, Rel. Min. Nancy Andrighi, j. em 23/08/2011.

in omittendo[103], caso o provedor deixasse de agir após ter sido comunicado de um conteúdo ilícito[104].

A regra fixada pelo Marco Civil da Internet, contudo, foi diferente do que vinha entendendo a jurisprudência do STJ. O artigo 19 do Marco Civil da Internet estabelece que:

> Com o intuito de assegurar a liberdade de expressão e impedir a censura, o provedor de aplicações de internet somente poderá ser responsabilizado civilmente por danos decorrentes de conteúdo gerado por terceiros se, após ordem judicial específica, não tomar as providências para, no âmbito e nos limites técnicos do seu serviço e dentro do prazo assinalado, tornar indisponível o conteúdo apontado como infringente, ressalvadas as disposições legais em contrário.

Percebe-se, de início, que a Lei privilegiou a liberdade de expressão ao mesmo tempo em que visa impedir a censura, e colocou a responsabilidade do provedor de aplicação de forma excepcional, apenas quando presentes determinados requisitos[105].

Por essa razão, vê-se que o provedor de aplicação não tem o dever de verificar previamente e impedir um conteúdo de ser postado por terceiro (o que configuraria censura) porque ele não será responsabilizado pelos danos este pode causar. O provedor de aplicação apenas será responsabilizado se, notificado judicialmente de um conteúdo ilícito, não o remover em prazo determinado, ou seja, por um ato omissivo, em caso de inércia *após receber ordem judicial específica*[106].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[103] "Em suma, pois, tem-se que os provedores de conteúdo: (i) não respondem objetivamente pela inserção no site, por terceiros, de informações ilegais; (ii) não podem ser obrigados a exercer um controle prévio do conteúdo das informações postadas no site por seus usuários; (iii) devem, assim que tiverem conhecimento inequívoco da existência de dados ilegais no site, removê-los imediatamente, sob pena de responderem pelos danos respectivos; (iv) devem manter um sistema minimamente eficaz de identificação de seus usuários, cuja efetividade será avaliada caso a caso". BRASIL. STJ, RESP 1.193.764/SP, 3ª Turma, Rel. Min. Nancy Andrighi, j. em 14/12/2010.

[104] Tal entendimento trouxe para a realidade brasileira a teoria do *notice and take down* americana. No entanto, a importação desse mecanismo ao Brasil foi feita sem um procedimento regulado, sem previsão de contranotificação e de outras garantias existentes nos Estados Unidos.

[105] Conforme afirma Caitlin Mulholland: "Parece claro que o legislador fez uma opção manifesta por privilegiar a liberdade de expressão e vedar qualquer tipo de censura prévia por parte do provedor de aplicação, ao excluir a responsabilidade civil do provedor a priori. Há verdadeiro posicionamento do legislador favoravelmente à livre manifestação de ideias (e contrariamente à censura) ao garantir que o provedor não será responsabilizado pela mera inclusão de conteúdo por terceiro em sua aplicação, ainda que este conteúdo seja considerado por um juízo a posteriori como ilícito, abusivo e violador de direitos". MULHOLLAND, Caitlin. Responsabilidade civil indireta dos provedores de serviço de Internet e sua regulação no Marco Civil da Internet. *In*: CELLA, José Renato Gaziero; NASCIMENTO, Aires Jose Rover, Valéria Ribas do. (orgs). *Direito e novas tecnologias*. 1ª Ed. Florianópolis: CONPEDI, 2015, v. 1, p. 485.

[106] Como sumarizam Carlos Affonso Souza e Chiara Spadaccini de Teffé: "Pode-se afirmar, portanto, que no artigo 19 do MCI: i) restou clara a responsabilidade subjetiva por omissão do provedor de aplicações de internet que não retira o conteúdo ofensivo após a devida notificação judicial; ii) como regra, a mera notificação extrajudicial não ensejará o dever jurídico de retirada do material questionado; iii) a opção de responsabilidade de viés subjetivo coaduna-se com o fim de

Por isso, é possível concluir que o Marco Civil da Internet definiu a responsabilidade subjetiva e solidária dos provedores de aplicação, afastando a tese de responsabilidade objetiva[107].

O parágrafo 1º do artigo 19 prevê que a ordem judicial "deverá conter, sob pena de nulidade, identificação clara e específica do conteúdo apontado como infringente, que permita a localização inequívoca do material". A jurisprudência entende tal indicação clara e específica do conteúdo como sendo a indicação da URL[108] de cada postagem[109].

O legislador previu uma exceção ao regime geral de responsabilização no parágrafo 2º do artigo 19, que dispõe que o *caput* não se aplica às infrações de direitos autorais ou conexos, hipótese que fica a depender de "previsão legal específica"[110]. Na prática da jurisprudência, acabou prevalecendo, para casos de infração de direitos autorais ou conexos, o entendimento anterior do STJ, de

PUC-Rio - Certificação Digital Nº 2112697/CA

assegurar a liberdade e evitar a censura privada na rede; iv) o Poder Judiciário foi considerado a instância legítima para definir a eventual ilicitude do conteúdo questionado e para construir limites para a expressão na rede, o que, por consequência, também promove uma maior segurança para os negócios desenvolvidos na Internet; e v) a remoção de conteúdo não dependerá exclusivamente de ordem judicial, de forma que o provedor poderá, a qualquer momento, optar por retirar o conteúdo caso ele seja contrário aos termos de uso de sua plataforma." SOUZA, Carlos Affonso Souza; TEFFÉ, Chiara Spadaccini de. *Responsabilidade dos provedores por conteúdos de terceiros na internet*. Disponível em: <https://www.conjur.com.br/2017-jan-23/responsabilidade-provedor-conteudo-terceiro-internet>. Acesso em: 8 Jan. 2022.

[107] Nesse sentido: "(...) a regra a ser utilizada para a resolução de uma dada controvérsia deve levar em consideração o momento de ocorrência do ato ou, em outras palavras, quando foram publicados os conteúdos infringentes. Para fatos ocorridos antes da entrada em vigor do Marco Civil da Internet, deve ser obedecida a jurisprudência desta corte. No entanto, após a entrada em vigor da Lei 12.965/2014, o termo inicial da responsabilidade solidária do provedor de aplicação, por força do art. 19 do Marco Civil da Internet, é o momento da notificação judicial que ordena a retirada de determinado conteúdo da internet". BRASIL. STJ, REsp 1.642.997, 3ª Turma, Rel. Min. Nancy Andrighi, j. em 12/09/2017.

[108] "A URL é o endereço de qualquer site na internet, mas pouca gente sabe como ela funciona. Há uma razão para cada um dos termos que muita gente acha esquisitos, que continuam a ser usados (por enquanto) porque simplesmente funcionam muito bem." Disponível em: <https://tecnoblog.net/responde/o-que-e-ur/>. Acesso em: 8 Jan. 2022.

[109] "1. A jurisprudência do STJ, em harmonia com o art. 19, § 1º, da Lei nº 12.965/2014 (Marco Civil da Internet), entende ser necessária a notificação judicial ao provedor de conteúdo ou de hospedagem para retirada de material ali publicado por terceiros usuários e apontado como infringente à honra ou à imagem dos eventuais interessados, sendo imprescindível a indicação clara e específica da URL - Universal Resource Locator - correspondente ao material que se pretenda remover". BRASIL. STJ, Agint no Agint no AREsp 956.396/MG, Rel. Min Ricardo Villas Bôas Cueva, j. em 17/10/2017.

[110] "§ 2º A aplicação do disposto neste artigo para infrações a direitos de autor ou a direitos conexos depende de previsão legal específica, que deverá respeitar a liberdade de expressão e demais garantias previstas no art. 5º da Constituição Federal".

responsabilidade *subjetiva* a partir de ciência do ilícito, *sem* a necessidade de ordem judicial específica[111].

Além dessa exceção, o artigo 21 da Lei também prevê que o regime de responsabilidade do artigo 19, *caput*, não se aplica em casos de veiculação de material contendo "cenas de nudez ou de atos sexuais de caráter privado". A veiculação desse tipo de material ficou conhecida pelo termo pornografia de vingança, ou *revenge porn*, definida como "uma espécie de exposição não autorizada de imagem íntima, em razão de suas características e dos sujeitos envolvidos"[112]. A exceção do artigo 21 se justifica para que nesses casos a remoção seja mais rápida, diante da gravidade da ofensa[113]. Assim, em caso de veiculação de material de pornografia de vingança[114], permanece o entendimento pela responsabilidade civil *subjetiva* do provedor de aplicação a partir da inércia após a ciência *extrajudicial* do conteúdo ilícito. A lógica por trás desse dispositivo é a presunção de que a "ilicitude de conteúdo de nudez é objetivamente verificável, ao contrário do que ocorre em casos de violações à honra, privacidade, intimidade,

PUC-Rio - Certificação Digital Nº 2112697/CA

[111] Tal ponto é criticado por Guilherme Magalhães Martins, que afirma que tal exceção demonstra que o Marco Civil da Internet deu prevalência para "situações jurídicas patrimoniais e dos fatores reais de poder sobre as existenciais, sobretudo das partes mais débeis. (…) Logo, o patrimônio, para o marco civil, prevalece sobre a cláusula geral de proteção da pessoa humana. Se a responsabilidade do provedor em face das vítimas depende de uma prévia notificação judicial, isso não se aplica, portanto, ao titular do direito autoral. Conferir aos interesses da indústria cultural, em função da titularidade dos direitos patrimoniais do autor (copyright) em face das vítimas de danos sofridos por meio das ferramentas de comunicação da internet, como as redes sociais, significa inverter os valores fundamentais contidos na tábua axiológica da Constituição da República. A vaga referência à futura Lei de Direitos Autorais, em discussão há mais de dez anos, não resolve o problema". MARTINS, Guilherme Magalhães. *Responsabilidade objetiva do provedor de aplicações de internet*. Disponível em: <https://www.conjur.com.br/2015-nov-18/guilherme-martins-responsabilidade-objetiva-provedor-internet>. Acesso em: 23 nov. 2022.

[112] TEFFÉ, Chiara Spadaccini de. Exposição não consentida de imagens íntimas: como o direito pode proteger as mulheres? *In:* ROSENVALD, Nelson; DRESCH, Rafael de Freitas Valle; WESENDONCK, Tula (coord.). *Responsabilidade civil:* novos riscos. Indaiatuba: Foco, 2019. p. 94.

[113] "Art. 21. O provedor de aplicações de internet que disponibilize conteúdo gerado por terceiros será responsabilizado subsidiariamente pela violação da intimidade decorrente da divulgação, sem autorização de seus participantes, de imagens, de vídeos ou de outros materiais contendo cenas de nudez ou de atos sexuais de caráter privado quando, após o recebimento de notificação pelo participante ou seu representante legal, deixar de promover, de forma diligente, no âmbito e nos limites técnicos do seu serviço, a indisponibilização desse conteúdo.

Parágrafo único. A notificação prevista no caput deverá conter, sob pena de nulidade, elementos que permitam a identificação específica do material apontado como violador da intimidade do participante e a verificação da legitimidade para apresentação do pedido".

[114] "É necessário ressaltar que o artigo 21 do Marco Civil engloba a situação chamada de pornografia de vingança, mas não somente ela, visto que o legislador não fez referência à motivação do agente." TEFFÉ, Chiara Spadaccini de. op. cit., p. 108.

imagem, ou mesmo desinformação, discurso de ódio e ataques antidemocráticos"[115].

Embora não tenha sido explícito – como é a Seção 230 do CDA –, o Marco Civil da Internet não contém nenhuma proibição para que as plataformas digitais moderem conteúdo com base em suas regras internas aceitas pelos usuários ao ingressarem nas respectivas plataformas. Entende-se, portanto, que plataformas digitais podem (e devem, para permitir um ambiente *online* saudável) moderar o conteúdo postado pelos usuários com base em seus termos de uso e diretrizes de comunidade[116]. Assim esclarecem Ana Frazão e Ana Rafaela Medeiros:

> Com efeito, a partir do Marco Civil da Internet - pelo menos se interpretado literalmente o art. 19 - os provedores passaram a gozar de uma dupla vantagem: (i) a faculdade de proceder unilateralmente, à moderação de conteúdo, de acordo com seus interesses e (ii) a ausência de quaisquer deveres de cuidado perante seus usuários por danos de conteúdo de terceiros, cabendo-lhes tão somente o cumprimento de ordem judicial específica[117].

No Brasil, contudo, a última palavra sobre qual conteúdo fica ou sai da rede é do Poder Judiciário[118]. Nos Estados Unidos, como visto no Capítulo 1.3.1 acima, as plataformas têm ampla liberdade para remover conteúdo de acordo com suas regras privadas, e ações no judiciário pedindo o reestabelecimento de um conteúdo não são concedidas[119].

Vale destacar que a constitucionalidade do artigo 19, do Marco Civil da Internet, está atualmente em debate perante o Supremo Tribunal Federal ("STF"),

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[115] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 150.

[116] Afinal, como sumarizam Chiara Spadaccini de Teffé e Carlos Affonso Souza: "A missão da Lei foi a de encontrar um equilíbrio entre a criação de um espaço onde fosse possível cultivar as liberdades de expressão e de informação e, ao mesmo tempo, garantir à vítima da disponibilização de conteúdo lesivo os meios adequados para identificar seu ofensor e para remover o material impugnado. De um lado, o MCI retira do provedor a pressão de remover todo e qualquer conteúdo apontado como ilícito, o que atingiria em cheio a liberdade de expressão, mas, de outro, não impede que assim ele proceda caso entenda que o material questionado é realmente contrário aos termos de uso e demais políticas que regem o funcionamento de sua plataforma." TEFFÉ, Chiara Spadaccini de.; SOUZA, Carlos Affonso. *Responsabilidade civil de provedores na rede: análise da aplicação do marco civil da internet pelo superior tribunal de justiça*. Revista IBERC, Minas Gerais, v.1, n.1, p. 01-28, nov.-fev./2019.

[117] FRAZÃO, Ana; MEDEIROS, Ana Rafaela. Responsabilidade civil dos provedores de internet: A liberdade de expressão e o art. 19 do Marco Civil. *In*: JÚNIOR; Marcos Ehrhardt; LOBO, Fabíola Albuquerque; e ANDRADE, Gustavo. *Liberdade de expressão e relações privadas*. Editora Fórum, *e-book*, p. 421.

[118] AFFONSO SOUZA, Carlos. *Bolsonaro edita decreto para acelerar liberação de emendas às vésperas da eleição*. Disponível em: <https://www1.folha.uol.com.br/mercado/2022/09/bolsonaro-edita-decreto-para-acelerar-liberacao-de-emendas-as-vesperas-da-eleicao.shtml>. Acesso em: 12 out. 2022.

[119] ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação*. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021.

com a repercussão geral dada ao Recurso Extraordinário 1.037.396/SP, interposto pelo Facebook contra acórdão da Turma Recursal de São Paulo, que responsabilizou a empresa por conteúdo postado por terceiro em sua plataforma, sem que a empresa tivesse violado qualquer ordem judicial pela remoção do conteúdo, como exige o artigo 19 do Marco Civil da Internet. O STF, ao reconhecer a repercussão geral, apontou que a discussão contrapõe "a dignidade da pessoa humana e a proteção aos direitos da personalidade à liberdade de expressão, à livre manifestação do pensamento, ao livre acesso à informação e à reserva de jurisdição"[120].

A doutrina contrária ao artigo 19, à qual se filiam Anderson Schreiber[121], Marcelo Calixto[122], Ana Frazão[123], Bruno Miragem[124], João Quinelato Queiroz[125], Guilherme Martins[126] e outros, entende que o regime escolhido pela legislação representa um retrocesso, pois deu menos proteção às vítimas do que a jurisprudência já vinha conferindo. Essa corrente entende, em resumo, que o artigo 19 tem quatro grandes equívocos.

Em primeiro lugar, assumiu a liberdade de expressão como princípio constitucional com posição privilegiada no ordenamento, sendo que os direitos fundamentais da pessoa humana (honra, privacidade, imagem, entre outros) também são tutelados pela Constituição brasileira em patamar axiológico não inferior à liberdade de expressão. Em segundo lugar, criticam a necessidade de

PUC-Rio - Certificação Digital Nº 2112697/CA

120 STF. Plenário Virtual. <https://portal.stf.jus.br/jurisprudenciaRepercussao/verPronunciamento.asp?pronunciamento=7397505>. Acesso em: 10 nov. 2022.

121 SCHREIBER, Anderson. *Marco Civil da Internet: avanço ou retrocesso*? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro. Disponível em: <andersonschreiber.com.br/downloads/artigo-marco-civil-internet.pdf>. Acesso em : 8 Jan. 2022.

122 CALIXTO, Marcelo Junqueira. Desindexação total e parcial nos motores de busca. *In*. SCHREIBER, Anderson, et al. *Direitos fundamentais e sociedade tecnológica*. Indaiatuba, SP: Editora Foco, 2022.

123 FRAZÃO, Ana; MEDEIROS, Ana Rafaela. Responsabilidade civil dos provedores de internet: A liberdade de expressão e o art. 19 do Marco Civil. *In*: JÚNIOR; Marcos Ehrhardt; LOBO, Fabíola Albuquerque; e ANDRADE, Gustavo. *Liberdade de expressão e relações privadas*. Editora Fórum, *e-book*.

124 MIRAGEM, Bruno. *Responsabilidade por danos na sociedade de informação e proteção do consumidor: desafios atuais na regulação jurídica da internet*. Revista de Direito do Consumidor, São Paulo, v. 70, p. 1-42, abr. 2009. Disponível em: <https://bdjur.stj.jus.br/jspui/handle/2011/83827>. Acesso em: 24 jun. 2022.

125 QUINELATO, João de Queiroz. *Responsabilidade Civil na Rede: danos e liberdades à luz do Marco Civil da Internet*. 1. ed. Rio de Janeiro: Processo, 2019. v. 1.

126 MARTINS, Guilherme Magalhães. *Vulnerabilidade e responsabilidade civil na Internet: a inconstitucionalidade do Artigo 19 do Marco Civil*. Revista de Direito do Consumidor , v. 137, p. 33-59, 2021.

ordem judicial para remover conteúdo ilícito ou ofensivo por criar um ônus à vítima, que terá que contar com a lentidão e custos do judiciário, em um tipo de dano que corre em velocidade na rede[127][128]. Terceiro, entendem que há discricionariedade técnica dos provedores para determinar a possibilidade ou não do cumprimento de uma ordem judicial. Por fim, consideram que a exigência de indicação clara e específica do conteúdo infringente (artigo 19, §1º) é mais um ônus às vítimas e um entrave prático à rapidez na apreciação do pedido e na multiplicação do conteúdo na rede[129].

Por outro lado, a doutrina a favor do regime de responsabilidade civil adotado pelo Marco Civil da Internet, com a qual se concorda, entende que a Lei agiu bem ao dar posição de destaque para a liberdade de expressão, e que o Poder Judiciário é a instância adequada para dirimir eventuais divergências sobre a ilicitude de um conteúdo. Com efeito, a ordem judicial é necessária e traz segurança às partes no caso concreto, porque permite uma avaliação judicial prévia sobre a potencial violação de direitos que necessitam de proteção jurídica. É o que entende Caitlin Mulholland:

> A necessidade da notificação judicial como requisito essencial para a responsabilização do provedor é uma medida, portanto, necessária e que traz segurança às partes envolvidas no caso concreto, pois permite uma avaliação judicial prévia sobre a potencial violação de direitos que necessitam de proteção jurídica. Ainda que de forma preliminar, a notificação judicial é procedimento judicializado e, portanto, requer a análise por um juiz por meio de um devido processo legal[130].

Afinal, o sistema de *notice and take down* pode levar a abuso de direitos porque permite, com simples notificação, que provedores removam diretamente o

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[127] SCHREIBER, Anderson. *Marco Civil da Internet: avanço ou retrocesso*? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro. Disponível em: <andersonschreiber.com.br/downloads/artigo-marco-civil-internet.pdf>. Acesso em : 8 Jan. 2022.

[128] QUEIROZ, João Quinelato de. op. cit., p. 129.

[129] Nesse sentido também entende João Victor Rozatti Longhi: "Não obstante, no caso de danos à personalidade perpetrados pela Rede, é comum que as informações se multipliquem rapidamente. Quando o usuário efetua o pedido para a retirada indica URLs que encontra e que estão naquele momento na Rede mundial de computadores. Por essa razão, já decidiu o STJ que incumbe a quem administra o site o dever técnico de impedir a divulgação do conteúdo ilícito, não lhe impondo a tarefa hercúlea de indicar precisamente as URLs (…)". LONGHI, João Victor Rozatti. *Marco Civil da Internet no Brasil:* Breves considerações sobre seus fundamentos, princípios e análise crítica do regime de responsabilidade dos provedores. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4635703/mod_resource/content/1/capi%CC%81tulo%205%20DIREITO%20PRIVADO%20E%20INTERNET.pdf>. Acesso em: 23 nov. 2022.

[130] MULHOLLAND, Caitlin. Responsabilidade civil indireta dos provedores de serviço de Internet e sua regulação no Marco Civil da Internet. *In*: CELLA, José Renato Gaziero; NASCIMENTO, Aires Jose Rover, Valéria Ribas do. (orgs). *Direito e novas tecnologias*. 1ª Ed. Florianópolis: CONPEDI, 2015, v. 1, p. 495.

conteúdo mediante critérios subjetivos além daqueles constantes em seus termos de uso, podendo afetar a liberdade de expressão. Como esclarecem Chiara Spadaccini de Teffé e Carlos Affonso Souza ao analisar o artigo 19, do Marco Civil da Internet:

> Entende-se que essa norma traz equilíbrio e proporcionalidade ao regime de responsabilidade por conteúdo de terceiro na Internet, bem como segurança jurídica e proteção aos novos modelos de negócio no País. Na ausência de uma adequada responsabilização, teríamos que enfrentar consequências negativas como, por exemplo, a diminuição da confiança de usuários e intermediários no uso e no desenvolvimento de ferramentas de comunicação da Internet, como o estímulo de ações governamentais e de agentes privados a estabelecerem mecanismos de controle e censura na Internet, o que levaria a processos arbitrários de remoção de conteúdo e vigilância dos cidadãos[131].

A necessidade de ordem judicial estimula provedores a não remover um conteúdo – que esteja em conformidade com seus termos de uso e políticas internas – apenas porque o mesmo gerou uma notificação, e incentiva assim que a vítima busque o Poder Judiciário e fundamente, em sede de ação judicial, os motivos pelos quais determinado conteúdo precisa ser removido. Do contrário, os próprios provedores estariam autorizados a decidir se um conteúdo impugnado, que não viole as políticas internas da plataforma, causa ou não um dano e se pode ou não ser exibido, o que por certo contaria com critérios subjetivos, a prejudicar a diversidade e o grau de inovação na internet, e podendo constituir censura privada. Além disso, poderia implicar entrave para o desenvolvimento de novas alternativas de exploração e comunicação, que poderiam não ser desenvolvidas em razão do receio de futuras ações compensatórias[132].

Conclui-se, portanto, que o Marco Civil da Internet agiu bem ao limitar a responsabilidade civil do provedor de aplicação ao descumprimento de ordem judicial específica, pois o Poder Judiciário é a instância mais apropriada para apreciar conteúdos desafiados, garantindo-se, assim, maior segurança às relações desenvolvidas *online*[133].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[131] TEFFÉ, Chiara Spadaccini de.; SOUZA, Carlos Affonso. *Responsabilidade civil de provedores na rede: análise da aplicação do marco civil da internet pelo superior tribunal de justiça.* Revista IBERC, Minas Gerais, v.1, n.1, p. 01-28, nov.-fev./2019.

[132] Ibid.

[133] "Não se pode exigir dos provedores que determinem o que é ou não apropriado para divulgação pública. Cabe ao Poder Judiciário, quando instigado, aferir se determinada manifestação deve ou não ser extirpada da rede mundial de computadores e, se for o caso, fixar a reparação civil cabível contra o real responsável pelo ato ilícito. Ao provedor não compete avaliar eventuais ofensas, em virtude da inescapável subjetividade envolvida na análise de cada caso. Somente o descumprimento de uma ordem judicial, determinando a retirada específica do material ofensivo, pode ensejar a reparação civil. Para emitir ordem do gênero, o Judiciário avalia a ilicitude e a repercussão na vida do ofendido no caso concreto. Ademais, mesmo não sendo aplicável ao caso, pois os fatos narrados nos autos são anteriores à sua vigência, observa-se que o Marco Civil da Internet (Lei nº

1.4 As diferentes formas utilizadas por plataformas de redes sociais para moderar conteúdo

Como visto, as plataformas de redes sociais exercem hoje grande influência sobre o discurso público e a liberdade de expressão. Diante da nova disposição "triangular" da liberdade de expressão, com as redes sociais em uma das pontas de controle[134], é necessário compreender as formas pelas quais tais empresas praticam a moderação de conteúdo postado por terceiros.

O presente capítulo pretende apresentar as principais formas de moderação de conteúdo utilizadas pelas redes sociais, que caracterizam a regulação do discurso público exercido por essas empresas privadas. Essas formas representam as possibilidades e os desafios enfrentados pelas plataformas ao lidar com uma escala massiva e sem precedentes de conteúdo de diferentes contextos em seus ambientes digitais. Para compreender os desafios envolvidos por trás da atividade de moderação de conteúdo, é preciso compreender não só as regras substantivas que cada empresa tem sobre o discurso, mas também o contexto tecnológico no qual o discurso é produzido e as possibilidades em que a moderação de conteúdo pode ocorrer[135].

Lawrence Lessig há anos já afirmou que a governança da internet se sustenta sobre quatro diferentes modais reguladores: (i) o direito, (ii) o mercado, (iii) as normas sociais e (iv) o código. Ou seja: na internet, o código é também espécie de lei[136]:

> Do mesmo modo que um protesto de rua é configurado pelo espaço físico de praças ou avenidas, a arquitetura tecnológica (código) determina as condições de funcionamento da internet; determina o que é possível fazer online, quais são as

PUC-Rio - Certificação Digital Nº 2112697/CA

12.965/2014) disciplinou, em seu artigo 19, o tema no sentido acima exposto (...) Em harmonia com os preceitos dessa norma, a jurisprudência do Superior Tribunal de Justiça tem entendido que a responsabilidade dos provedores de hospedagem e de conteúdo depende da indicação, pelo autor, do respectivo URL (Universal Resource Locator) em que se encontra o material de cunho impróprio". STJ, REsp, 1.568.935, 3ª Turma, Rel. Min. Ricardo Villas Bôas Cueva, j. em 05/04/2016.

[134] BALKIN, Jack M. *Old-school/New-school speech regulation*. Harvard Law Review, Forthcoming, Yale Law School, Public Law Research Paper No. 491, 2014. Disponível em: <https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2888582_code293225.pdf?abstractid=2377526&mirid=1&type=2>. Acesso em: 04 abr. 2022.

[135] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 48.

[136] LESSIG, Lawrence. *Code: version 2.0*, Basic Books, 2006, p. 122/137.

> 'leis da física' daquele ambiente. Como códigos nunca são neutros, é necessário estar atento às escolhas e às possibilidades de controle que neles estão embutidas[137].

Um exemplo de como os códigos importam para a governança de discursos na internet pode ser visto nos algoritmos, que não só determinam o que vai ter mais ou menos visibilidade para determinado usuário, como também criam incentivos para que as pessoas produzam conteúdos que sejam valorizados por esse código.

Como se verá adiante, arquiteturas tecnológicas e procedimentais (código) construídas pelas plataformas de redes sociais são relevantes para entender como é feita a moderação de conteúdo em seus ambientes. Existem hoje novas formas de regular discursos que são praticadas pelas plataformas digitais, em comparação às formas tradicionais da "velha escola"[138].

1.4.1 Controle automatizado de conteúdo feito de forma prévia

O controle prévio é a mais restritiva forma de moderação de conteúdo, pois implica na ideia de controle feito antes de o conteúdo ser disponibilizado, com base nas regras previamente disponibilizadas aos usuários. O controle prévio automatizado é importante pois constitui exceção à regra geral de que tudo pode ser postado em plataformas de redes sociais sem qualquer filtragem prévia.

Essa forma de moderação de conteúdo é vista regularmente nos casos de *uploads* de vídeos no Facebook e Youtube, por exemplo. É uma forma de controle prévio pois é feita no momento entre o envio do vídeo que se pretende postar, e sua efetiva publicação. Nesse momento no Facebook, por exemplo, há a mensagem: "Processando o Vídeo: o vídeo da sua publicação está sendo processado. Nós iremos enviar uma notificação quando estiver pronto e seu vídeo apto a visualização". Trata-se, em regra, de um controle automatizado, efetuado por meio de filtragem algorítmica, sem revisão humana[139].

É uma importante forma de moderação de conteúdo para melhor construção de uma comunidade digital pois, apesar de parecer rígida, é geralmente utilizada

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[137] NITRINI, Rodrigo Vidal. op. cit.

[138] BALKIN, Jack M. *Old-school/New-school speech regulation*. Harvard Law Review, Forthcoming, Yale Law School, Public Law Research Paper No. 491, 2014. Disponível em: <https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2888582_code293225.pdf?abstractid=2377526&mirid=1&type=2>. Acesso em: 02 abr. 2022.

[139] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598, p. 1636.

em casos extremos (como pornografia), em que não há zona cinzenta sobre a ilicitude do conteúdo, evitando que casos de claro abuso sejam publicados.

O processo de desenvolvimento desse tipo de tecnologia se deu, principalmente, a partir de restrições às imunidades previstas na Seção 230, do CDA, nos Estados Unidos, no que tange a direitos autorais e pornografia infantil, se expandindo posteriormente também para casos de propaganda terrorista, esta última por decisão das próprias plataformas[140].

A legislação federal americana determina que empresas reportem a existência de pornografia infantil ao *International Center for Missing and Exploited Children* apenas quando tenham ciência da existência desse tipo de conteúdo em suas plataformas. Isso significa que plataformas não são obrigadas a monitorar, de forma proativa, a existência desse tipo de conteúdo. Mesmo assim, esse monitoramento prévio automatizado se tornou comum entre as plataformas como uma forma de responder proativamente uma demanda social[141].

Esse tipo de monitoramento permite a checagem sobre se uma determinada imagem corresponde ou não a um material previamente caracterizado como pornografia infantil por meio do algoritmo de reconhecimento de imagem chamado

PUC-Rio - Certificação Digital Nº 2112697/CA

[140] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 50.
[141] KOSSEF, Jeff. *The Twenty-Six Words That Created the Internet*. Cornell University Press, 2019, capítulo 12.

PhotoDNA[142], que usa como base vasto banco de dados do governo federal americano com cerca de 720 mil imagens ilegais existentes na internet[143].

Esse tipo de controle automatizado prévio em casos de pornografia infantil é relativamente simples e não gera maiores controvérsias, já que é um conteúdo objetivamente identificável como ilícito em todos os países. Essa tecnologia, contudo, avançou para utilização também no campo do direito autoral, após forte pressão da indústria para garantir que plataformas não se tornassem uma ameaça para suas atividades[144], e também por ser uma exceção à regra geral da Seção 230, do CDA. Em casos de disputa de direito autoral, as plataformas só fazem jus à imunidade da Seção 230 caso removam o conteúdo infrator de forma célere após terem conhecimento do mesmo (regra do *notice and take down*).

Dessa forma, o Youtube, por exemplo, desenvolveu, por cerca de R$ 60 milhões de dólares, o software *Content ID*[145], que permite a inclusão de materiais em um banco de dados, que passam a contar com uma *digital fingerprint*, espécie de marca identificadora digital, para verificar os vídeos postados na plataforma. A

PUC-Rio - Certificação Digital Nº 2112697/CA

[142] "O PhotoDNA usa uma técnica chamada "hashing", na qual uma imagem digital é transformada em uma sequência numérica com base na sequência de cores nos pixels individuais da imagem. Essa *hash* serve como um identificador, uma espécie de impressão digital, pois é única para cada imagem e é identificável nas cópias dessa imagem, mesmo que tenham sido alteradas em algum grau. Uma plataforma pode então pegar todas as imagens carregadas pelos usuários, comparar o *hash* exclusivo de cada uma com um banco de dados existente de imagens conhecidas de pornografia infantil mantidas pelo Centro Nacional para Crianças Desaparecidas e Exploradas (NCMEC). Se a imagem for compatível, a plataforma pode removê-la e alertar o NCMEC e possivelmente as autoridades. Cada foto que você publica no Instagram, cada imagem que você fixa no Pinterest, cada *snap* no Snapchat é rápida e automaticamente comparada com o banco de dados do NCMEC para garantir que não seja pornografia infantil".
Tradução livre de: "PhotoDNA uses a technique called "hashing," in which a digital image is turned into a numerical string based on the sequence of colors in the image's individual pixels. This string serves as an identifier, a kind of fingerprint, as it is unique to each image, and is identifiable in copies of that image, even if they've been altered to some degree. A platform can then take all images uploaded by users, compare the unique hash of each to an existing database of known child pornography images maintained by the National Center for Missing and Exploited Children (NCMEC). If the image is a match, the platform can remove it, and can alert NCMEC and possibly the authorities. Every single photo you post to Instagram, every image you pin on Pinterest, every snap on Snapchat, is quickly and automatically compared to the NCMEC database to make sure it is not child pornography". GILLESPIE, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven, CT: Yale University Press, 2018), p. 100. <https://unglueit-files.s3.amazonaws.com/ebf/5f82765552144327afd531625486f0e3.pdf>. Acesso em: 12 nov. 2022.

[143] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598, p. 1637.

[144] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 51.

[145] Disponível em: <https://support.google.com/youtube/answer/2797370?hl=en>. Acesso em: 12 nov. 2022.

automatização desse controle (que não impede que vídeos sejam marcados como violadores por detentores de direito autoral, utilizando-se o sistema de *notice and take down*) foi a forma encontrada pelo Youtube para manter uma parceria com a indústria cultural. Ao identificar um vídeo como coberto por direito autoral, a plataforma diretamente oferece aos detentores do direito as opções de bloquear o vídeo, aguardar para ver se ele se torna popular ou incluir propagandas que lhes tragam retorno direto[146]. Tem-se, assim, mais um exemplo de controle automatizado incentivado inicialmente pela legislação, mas desenvolvido e aprimorado pelas próprias necessidades do mercado, em uma espécie de autorregulação que foi posteriormente expandida para outras searas, como no caso de combate ao terrorismo:

> Em ambos os casos, imperativos de normas sociais e de mercado tiveram maior peso para o desenvolvimento do 'código'. Mas uma vez que essa nova possibilidade técnica estava cada vez mais avançada e consolidada, ela poderia ampliada para potencialmente qualquer nova seara. Inclusive por decisões autônomas das próprias plataformas, por uma espécie de autorregulação. Foi isso o que ocorreu - desta vez, na área de combate ao terrorismo[147].

Com efeito, em 2016, Facebook, Microsoft, Twitter e Youtube anunciaram que passariam a coordenar ações no combate ao terrorismo *online*. Isso se deu através do compartilhamento de bases de dados de *hashes* (como marcas identificadoras digitais) sobre imagens terroristas violentas ou vídeos de recrutamento já previamente removidos de alguma plataforma. Assim, cada empresa poderia adicionar novos conteúdos a esse banco de dados, que seria posteriormente utilizado pelas outras de modo a fazerem seu próprio controle sobre a remoção ou não do conteúdo, preservando a decisão final de cada plataforma[148].

Essa iniciativa chama atenção para a possibilidade e capacidade técnica que as plataformas digitais têm, sem qualquer iniciativa governamental, de formar uma base de dados concentrada e compartilhada para servir de parâmetro para o controle de conteúdo.

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[146] NITRINI, Rodrigo Vidal. Op. cit., p. 52.

[147] Ibid, p. 53.

[148] O comunicado conjunto das empresas explica que "cada empresa continuará a aplicar suas práticas de transparência e revisão para pedidos governamentais, além de manter seu próprio espaço para recursos para decisões de remoção e demais informações". Partnering to help curb the spread of terrorist content online, publicado em 05/12/2016. Disponível em: <https://blog.google/around-the-globe/google-europe/partnering-help-curb-spread-terrorist-content-online/>. Acesso em: 12 nov. 2022.

O Facebook, por exemplo, trabalha para avançar na possibilidade de filtragens prévias de vídeos ao vivo, em resposta aos episódios de suicídios sendo transmitidos ao vivo por usuários na plataforma. Atualmente, transmissões ao vivo são controladas por meio de denúncias de outros usuários. Esse tipo de controle, contudo, é mais difícil quando se lida com transmissões ao vivo, embora permita marcar um determinado vídeo e impedir sua republicação[149].

Isso foi feito pelo Facebook no caso do atentado de Christchurch, ocorrido em 2019 na Nova Zelândia, que resultou em 51 mortes. O atentado foi transmitido ao vivo no Facebook pelo atirador. Com a repercussão do caso, o Facebook iniciou seu protocolo de crise, que inclui equipes de plantão especializadas a qualquer momento, em todo o mundo. Depois de algumas horas, o procedimento adotado foi, após apagar a postagem original, marcar a identidade digital do vídeo para evitar sua repostagem. No entanto, o Facebook encontrou dificuldades nisso, pois usuários passaram a manipular digitalmente o vídeo para driblar a restrição imposta pela empresa. Em resposta, o Facebook passou a fazer uma identidade digital a partir da trilha de áudio, o que deu certo. De fato, 1,5 milhões de cópias do vídeo foram removidas do Facebook, sendo que 1,2 milhões foram removidas no momento do upload[150].

1.4.2 Análise automatizada de linguagem após a publicação de um conteúdo

Diante da enorme quantidade de conteúdo postado *online* a cada segundo, tornou-se necessário o desenvolvimento de alternativas à moderação efetuada por revisores humanos, como a análise automatizada de linguagem. O Facebook, por exemplo, em 2018, removeu quase metade do conteúdo relacionado a discurso de ódio de forma proativa, utilizando-se de tecnologia automatizada, incluindo o monitoramento de imagens e textos[151][152].

[149] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 56.
[150] KLONICK, Kate. *Inside the team at Facebook that dealt with the Christchurch shooting*. The New Yorker, 2016.
[151] Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2018-05/facebook-remove-25-milhoes-de-posts-com-discurso-de-odio-em-6-meses>. Acesso em: 24 nov. 2022.
[152] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 57.

PUC-Rio - Certificação Digital Nº 2112697/CA

Monica Bicket, *head of global policy management* do Facebook, já esclareceu que o uso desse tipo de ferramenta é mais eficaz para o monitoramento de imagens já previamente conhecidas, os *hashes*[153], mas ela tem sido mais utilizada para avaliação generalizada de conteúdo a partir da evolução de formas de aprendizado pela inteligência artificial[154].

Essa espécie de monitoramento, embora crescente e necessário, apresenta riscos e limitações na moderação de conteúdo. Afinal, tratam-se de plataformas globais, com conteúdo sendo gerado por todo o globo, e a inteligência artificial muitas vezes não está apta a compreender o contexto e o real sentido de uma postagem. Além disso, fato é que algoritmos que atuam por meio do *machine learning* incorporam vieses de quem cria o código, que podem ser discriminatórios[155].

Uma das ferramentas que utiliza esse tipo de monitoramento é a Perspective, de uma empresa de propriedade da Alphabet, controladora do Google[156]. Essa tecnologia mede o nível de toxidade em um texto, considerando tóxico um "comentário rude, desrespeitoso ou não razoável que provavelmente fará com que você abandone a discussão". O programa aprende e se aperfeiçoa com base nas avaliações de pessoas sobre se um conteúdo é 'muito saudável' ou 'muito tóxico'[157].

Um estudo conduzido pelo *Center for Democracy and Technology* ("CDT") apontou os riscos de vieses discriminatórios dessa ferramenta ao detectar, de forma automática, o que caracteriza discurso de ódio[158]. O estudo apontou que há resultados falso-positivo na ferramenta, que identificavam expressões comuns à população negra americana[159]. Em relação à identificação de discurso de ódio, o estudo apontou 5 limitações da ferramenta: (i) funciona melhor em contextos determinados, não havendo confiabilidade na ampliação de seu uso para outros contextos; (ii) possibilidade de reprodução de vieses discriminatórios contra grupos vulneráveis ou marginalizados; (iii) necessidade de se definir de forma acurada o

PUC-Rio - Certificação Digital Nº 2112697/CA

[153] Disponível em: <https://cyber.harvard.edu/events/state-online-speech-and-governance>. Acesso em: 12 nov. 2022.
[154] NITRINI, Rodrigo Vidal. Op. cit.
[155] Ibid, p. 58.
[156] Disponível em: <https://fortune.com/2017/02/23/alphabet-jigsaw-perspective-comment-moderator/>. Acesso em: 12 nov. 2022.
[157] NITRINI, Rodrigo Vidal. op. cit., p. 59.
[158] Disponível em: <https://aclanthology.org/P19-1163.pdf>. Acesso em: 12 nov. 2022.
[159] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 59.

conteúdo em questão, não podendo se limitar a 'extremismo'; (iv) baixa precisão de resultados - média entre 70% e 75% de acerto, sendo necessária aplicação da ferramenta de forma cautelosa e; (v) facilidade de evasão do monitoramento, mediante alterações de elementos contextuais[160].

1.4.3. *Flagging*

Flagging é o modo pelo qual os próprios usuários de uma plataforma marcam determinado conteúdo (*flag*) como sendo irregular ou indesejável, fazendo com que esse conteúdo passe por um processo de revisão, que geralmente é feito por moderadores humanos. Trata-se de uma denúncia a uma postagem que é realizada pelos próprios usuários.

Essa foi uma das formas adotada pelas plataformas para lidar com o desafio de escala das grandes redes sociais. Afinal, há um enorme volume de conteúdo sendo gerado a cada segundo, tornando impossível a revisão completa de todo conteúdo postado. Diante desse problema de escala, o *flagging* se tornou uma forma importante e necessária na atividade de moderação de conteúdo[161].

O uso desse tipo de moderação de conteúdo pelas redes sociais atende a duas principais funções: (i) é uma forma prática de revisar milhares de conteúdos; e (ii) o fato de se basear em denúncias dos próprios usuários serve para legitimar as decisões quando as plataformas são questionadas por censurar ou banir algum conteúdo[162].

Tarleton Gillespie esclarece que utilizar os próprios usuários como "fiscais" do conteúdo postado é conveniente para as plataformas, pois divide a complexa tarefa de moderar conteúdo entre muitos, além de conferir legitimidade às decisões das plataformas. Afinal, permitir que usuários possam *flag* conteúdo irregular ou indesejável sinaliza que elas estão ouvindo seus usuários e fornecendo caminhos para que eles possam expressar a ofensa e pedir ajuda quando se sentirem prejudicados[163].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[160] Ibid, p. 58.
[161] GILLESPIE, Tarleton. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 96.
[162] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598, p. 1638.
[163] GILLESPIE, Tarleton. op. cit.

Isso, conteúdo, também demanda altos investimentos das plataformas. O Facebook, por exemplo, recebe milhões de denúncias para revisar a cada dia. Ademais, há problemas inerentes em permitir que a própria "comunidade" – que na realidade não é uma comunidade, mas bilhões de usuários ao redor do mundo, em diferentes contextos – fiscalize o que está sendo postado. Na prática, muitas vezes grande parte das denúncias realizadas decorrem de disputas interpessoais entre usuários, e não propriamente por violação de políticas de conteúdo:

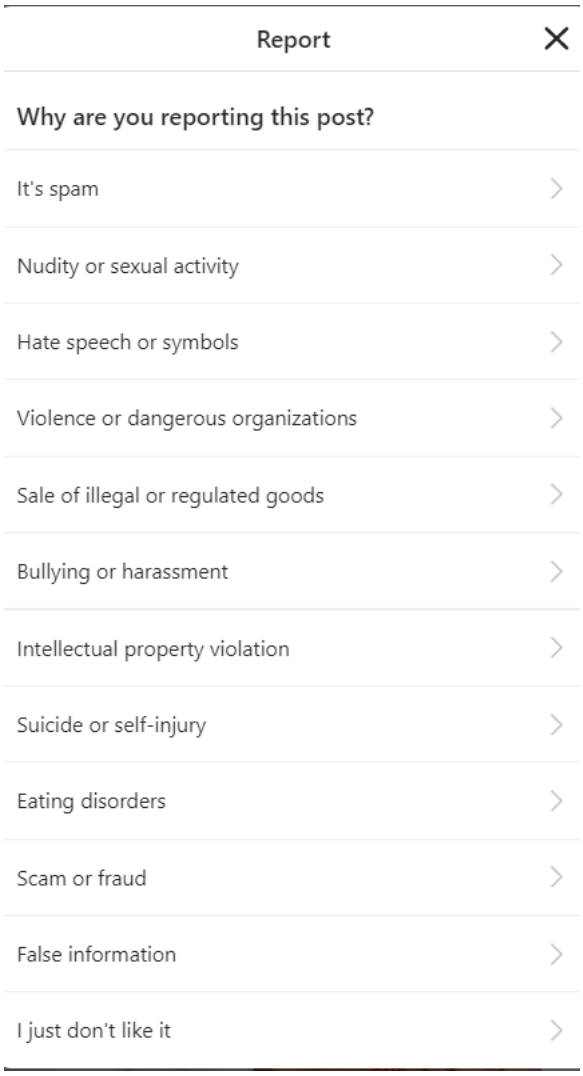
> Os sistemas de sinalização também podem ser manipulados, armados para realizar fins sociais e políticos. Há evidências de que a sinalização estratégica já ocorreu, e suspeita de que tenha ocorrido amplamente. Os usuários sinalizarão coisas que os ofendem politicamente, ou com os quais discordam; o fato de as políticas de moderação de uma plataforma ter sido violada pode ser irrelevante. A esperança é que muitos sinalizadores possam persuadir os moderadores da plataforma a removê-lo. Mesmo que uma plataforma seja diligente sobre reivindicar conteúdo sinalizado de forma inadequada, alguns conteúdos podem ainda ser removidos incorretamente, algumas contas podem ser suspensas[164].

Por essa razão, ao realizar uma denúncia, o usuário geralmente deve preencher um questionário, auxiliando as plataformas a organizarem as demandas e priorizar casos urgentes[165]. É o que ocorre, por exemplo, no Instagram:

PUC-Rio - Certificação Digital Nº 2112697/CA

[164] Tradução livre de: "Flagging systems can also be gamed, weaponized to accomplish social and political ends. There is evidence that strategic flagging has occurred, and suspicion that it has occurred widely. Users will flag things that offend them politically, or that they disagree with; whether a particular site guideline has been violated can be irrelevant. The hope is that enough flags might persuade platform moderators to remove it. Even if a platform is diligent about vindicating content that's flagged inappropriately, some content may still be removed incorrectly, some accounts may be suspended". GILLESPIE, Tarleton. *Custodians of the Internet:* Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, CT: Yale University Press, 2018, p. 101.

[165] Ibid, p. 99.



(janela de denúncia do Instagram, 2022).

Em todo caso, exemplo do mal uso do sistema de *flagging* pode ser visto em um caso ocorrido no Facebook, em 2014. O Facebook começou a suspender contas de *drag queens* por violação à política de obrigatoriedade de uso de nomes reais. Centenas de *drag queens* que criaram contas com seus nomes artísticos tiveram suas contas encerradas, ou receberam solicitações da plataforma para que informassem sua identificação "real". As pessoas afetadas, principalmente de São Francisco, e membros da comunidade LGBTQI+ que as apoiaram, protestaram e conversaram com representantes do Facebook. Após duas semanas de imprensa negativa, o Facebook cedeu, pedindo desculpas à comunidade e esclarecendo que o uso de nome artístico seria suficiente para manutenção das contas, alterando sua

PUC-Rio - Certificação Digital Nº 2112697/CA

interpretação sobre a regra do "nome real". No pedindo de desculpas, o Facebook informou que se tratou de ação coordenada de um único usuário[166][167].

1.5.4 Controle humano exercido por moderadores após a publicação de um conteúdo

A atividade de moderação de conteúdo das redes sociais, e especialmente a do Facebook, oferece importante destaque à moderação efetuada por revisores humanos, que realizam uma função necessária, pois, em muitos casos, apenas uma revisão humana permite que um conteúdo seja devidamente contextualizado e analisado diante das políticas da empresa. Esse tipo de moderação é geralmente feita *a posteriori*, depois que uma publicação é feita, sendo geralmente reativa, ou seja, feita após um conteúdo já ter sido reportado (*flagged*) por outros usuários[168].

Kate Klonick esclarece que, no Facebook, há três níveis de moderadores de conteúdo: (i) os moderadores do nível 3, que fazem a maior parte da revisão de conteúdo do dia a dia; (ii) moderadores nível 2, que fazem a supervisão do trabalho conduzido pelos moderadores do nível 3; e (iii) os moderadores do nível 1, que são normalmente advogados ou os próprios formuladores das políticas da empresa[169].

Atualmente, na maior parte das plataformas, incluindo no Facebook, os moderadores nível 3 trabalham em *call centers* de países subdesenvolvidos. No Facebook, eles são chamados de “equipes de suporte ao usuário”. Os moderadores de nível 3 normalmente revisam o material que foi sinalizado como menor prioridade pelo fluxo de denúncias. No Facebook, por exemplo, isso inclui, em parte, relatos de nudez ou pornografia, insultos ou ataques com base em religião,

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[166] GILLESPIE, Tarleton. *Custodians of the Internet:* Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, CT: Yale University Press, 2018, p. 103.
[167] “A forma como isso aconteceu nos pegou desprevenidos. Um indivíduo no Facebook decidiu denunciar várias centenas dessas contas como falsas. Esses relatórios estavam entre as várias centenas de milhares de relatórios de nomes falsos que processamos todas as semanas, 99% dos quais são pessoas mal-intencionadas fazendo coisas ruins: falsificação de identidade, intimidação, trollagem, violência doméstica, golpes, discurso de ódio e muito mais - então não observe o padrão.” Disponível em: <https://www.facebook.com/chris.cox/posts/10101301777354543>. Acesso em: 10 nov. 2022.
[168] KLONICK, Kate. T*he New Governors: The People, Rules, and Processes Governing Online Speech*, Harvard Law Review 131 (2018): 1598, p. 1638.
[169] Ibid, p. 1640.

etnia ou orientação sexual, conteúdo impróprio ou nojento, conteúdo humilhante ou conteúdo que defenda violência a uma pessoa ou animal[170].

Os moderadores nível 2 supervisam o trabalho dos de nível 3, e trabalham tanto de forma remota (a maioria vivendo nos Estados Unidos) como presencial, em *call centers*. Eles revisam conteúdo que foi priorizado, como ameaças iminentes, terrorismo e suicídio. Esse conteúdo chega ao nível 2 diretamente através do fluxo de denúncias ou sendo identificado e escalado para o nível 2 pelos moderadores do nível 3. Os moderadores de nível 2 também revisam algumas amostras aleatórias de decisões de moderação de nível 3. A fim de garantir a precisão da moderação, o Facebook e outras plataformas têm uma certa redundância embutida: o mesmo conteúdo é frequentemente dado a vários moderadores de nível 3. Se o julgamento sobre o conteúdo variar, o conteúdo é reavaliado por um moderador de nível 2[171].

A moderação do nível 1 é predominantemente realizada no departamento jurídico ou de políticas da sede da plataforma, na Califórnia. Atualmente, o Facebook tem cerca de 15 mil moderadores de conteúdo espalhados ao redor do mundo[172]. Esse modelo de moderação humana de conteúdo do Facebook permitiu que a plataforma pudesse escalar globalmente seu time de revisores, auxiliando no problema de escala da rede social, com conteúdo sendo postado em mais de 50 línguas diferentes. Isso requer que o Facebook contrate moderadores fluentes em diversas línguas para que sejam capazes não só de entender a linguagem, mas também o contexto em que cada publicação é feita e deve ser revisada[173].

De toda forma, não se trata de um trabalho fácil. Um dos principais problemas apontados nessa atividade é o efeito psicológico sobre a saúde mental dos moderadores, que são expostos, de forma regular, a conteúdos violentos e tóxicos, e são pouco remunerados por isso[174].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[170] Ibid.
[171] Ibid, p. 1641.
[172] Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2020/11/21/moderadores-de-conteudo-do-facebook-enviam-carta-aberta-a-zuckerberg.htm>. Acesso em: 12 nov. 2022.
[173] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 68.
[174] Parte dos moderadores nível 3, por exemplo, ganham por volta de 15 dólares por hora no Arizona. *The Trauma Floor: the secret lives of Facebook moderators in America*, The Verge, 25/02/2019.

1.5 Arcabouço histórico da moderação de conteúdo no Facebook: da aplicação de *standards* genéricos à construção de um sistema de regras

O poder concentrado nas plataformas de redes sociais hoje foi algo que ocorreu de forma inesperada para as próprias plataformas, e suas políticas de moderação de conteúdo foram aprimoradas e construídas com o tempo, na medida em que se tornaram cada vez mais necessárias. Em seus anos iniciais (anos 2000), as plataformas tinham uma política de moderação de conteúdo similar aquelas que ocorriam em fóruns de internet. Havia poucos responsáveis pela revisão do conteúdo publicado, que se baseavam em parâmetros genéricos e limitados[175]. As regras eram ocultas e opacas, e as plataformas decidiam livremente dentro do paradigma da autorregulação, amparadas pelas imunidades conferidas pela Seção 230, do CDA.

No caso do Facebook, quando Dave Willner, quem escreveu as regras sobre moderação de conteúdo da plataforma[176], ingressou como parte de um pequeno grupo especializado em moderação de conteúdo, em 2009, não havia "padrões de comunidade" na empresa. Naquele período, toda atividade de moderação era realizada com base em uma página de “regras” internas aplicadas globalmente para todos os usuários. Willner explica que:

> a política tinha cerca de uma página, uma lista de coisas que você deve deletar: então eram coisas como Hitler e pessoas peladas. Nenhuma dessas coisas era errada, mas não havia uma estrutura de regras para explicar por que aquelas coisas estavam na lista". "A regra era: se algo lhe faz mal, então apague[177].

No entanto, esse modelo foi ficando insustentável na medida em que a plataforma deixou de se limitar a uma comunidade mais homogênea de universitários americanos, tornando-se cada vez mais global e diversificada. Diante disso, o paradigma da autorregulação passou a ser questionado, e a legitimidade das plataformas para decidir o que pode ou não permanecer *online* entrou em crise,

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[175] KLONICK, Kate. “*The New Governors: The People, Rules, and Processes Governing Online Speech*,” Harvard Law Review 131 (2018): 1598, p. 1631.

[176] Disponível em: <https://www.wired.com/story/facebook-content-standards-politicians-exemption-dave-willner/>. Acesso em: 12 nov. 2022.

[177] Tradução livre de: "The [policy] guidance was about a page; a list of things you should delete: so it was things like Hitler and naked people. None of those things were wrong, but there was no explicit framework for why those things were on the list." “if it makes you feel bad in your gut, then go ahead and take it down.” KLONICK, Kate. op. cit., p. 1631.

iniciando movimentos para tentar limitar a arbitrariedade no exercício do poder dessas empresas. Diversos países passaram, por exemplo, a pressionar as plataformas para que fossem mais eficientes no combate à desinformação *online*[178].

Assim, no final de 2009, houve a formação de um grupo especializado de cerca de 12 moderadores no Facebook, com a tarefa de sistematizar o que se transformou na primeira versão das "regras de comunidade"[179].

Willner baseou seu primeiro conjunto de regras para o Facebook utilizando como modelo códigos contra assédio de sua universidade. Ele percebeu, conteúdo, que *standards* vagos, como a proibição de discursos que criem "ambiente hostil", levavam a dúvidas e decisões divergentes entre os moderadores, que naquele momento já estavam em diferentes contextos culturais ao redor do mundo. Diante das diferentes concepções sobre os temas da liberdade de expressão e discurso de ódio entre os Estados Unidos e a Europa, por exemplo, Willner passou a buscar uma política contra discursos de ódio "que focassem em ações concretas, facilmente identificáveis", para que a decisão de eventual remoção "pudesse ser baseada em nada além das informações contidas no formulário que usuários do Facebook utilizam para fazer uma reclamação sobre posts ofensivos, aplicadas como um algoritmo"[180]. Para o autor Jeffrey Rosen, Willner tentou oferecer uma resposta de engenheiro para um problema histórico e jurídico complexo, "uma abordagem ao estilo do Vale do Silício"[181].

Embora Willner tenha usado o termo "regras" para descrever as possibilidades de remoção de conteúdo, seria mais preciso o uso do termo *standards* nesse momento inicial. Kate Klonick traz um exemplo de *standard* como "não dirija muito rápido", enquanto uma regra nesse sentido seria trazer um limite de velocidade de 65 milhas por hora. Há pontos negativos e positivos em escolher regular por meio de regras ou *standards*. Estes são geralmente reafirmações de valores e princípios, mas são vagos e opacos, e assim são sujeitos a decisões arbitrárias. Por isso mesmo, contudo, podem ser aplicados de forma precisa e

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[178] Disponível em: <https://itsrio.org/wp-content/uploads/2021/04/Relatorio_RedesSociaisModeracaoDeConteudo.pdf>. Acesso em: 12 nov. 2022.

[179] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598, p. 1631.

[180] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 85.

[181] ROSEN, Jeffrey. "*The Deleted Squad*", artigo publicado pelo The New York Republic, em 29/04/2013.

eficiente, e podem acomodar melhor mudanças de circunstâncias. Regras, por outro lado, podem ser mais fáceis de serem cumpridas, mas podem ser excessivas e levar a resultados injustos. Regras permitem pouca discricionariedade e, nesse sentido, limitam os caprichos dos moderadores, mas também podem conter lacunas e conflitos, gerando complexidade e litígio[182].

As mudanças das políticas das plataformas de *standards* para regras refletem os ônus e bônus de cada um. Um *standard* genérico sobre algo como violência gratuita é capaz de alcançar uma decisão mais justa e precisa que reflita as normas da comunidade, mas é vaga, dependente de fatos, e mais custosa de ser aplicada. O Facebook, por exemplo, abandonou sua política com base em *standards* conforme o volume de conteúdo gerado por terceiros aumentou, a base de usuários se diversificou e os moderadores de conteúdo se espalharam pelo mundo[183].

Para Willner, era necessário eliminar *standards* que evocavam valores, sentimentos, e outras reações subjetivas em seu novo conjunto de regras sobre moderação. Por isso, ele focou em criar regras objetivas. A primeira versão das regras do Facebook tinha quinze mil palavras, tendo por principais objetivos manter a consistência e uniformidade para obter o mesmo julgamento sobre determinado conteúdo, independentemente de quem fosse o moderador responsável[184]. Essa, conteúdo, não é uma tarefa fácil. Ao tentar pegar para si a tarefa de criar um sistema de regras sobre discursos para aplicá-las de forma global em diferentes contextos, o Facebook acabou abraçando dilemas que há muito fazem parte da pauta de debates sobre liberdade de expressão[185].

Assim, a moderação de conteúdo no Facebook (e em outras plataformas, como o Youtube), se desenvolveu de um sistema de *standards* para um sistema de regras, por conta de 3 fatores, explicados por Kate Klonick: (i) o rápido crescimento do número de usuários e do volume de conteúdo gerado; (ii) a globalização e diversidade da comunidade; e (iii) aumento da dependência em equipes de moderadores humanos com diversas formações.

PUC-Rio - Certificação Digital Nº 2112697/CA

[182] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, Harvard Law Review 131 (2018): 1598, p. 1632.
[183] Ibid, p. 1633.
[184] Ibid, p. 1634.
[185] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 85.

Em abril de 2018, após a eleição do ex-Presidente Donald Trump com campanha baseada principalmente no uso de redes sociais, de materiais internos da plataforma serem vazados e publicados pelo *The Guardian*[186], e da edição de leis europeias responsabilizando redes sociais por conteúdo "manifestamente ilegal", como a NetzDG, na Alemanha (analisada no Capítulo 2.5.1), o Facebook publicou pela primeira vez as *guidelines* internas que usava para aplicar os Padrões da Comunidade[187].

Em novembro de 2018, o fundador e CEO do Facebook, Mark Zuckerberg, falou sobre criar um órgão independente para avaliar a moderação de conteúdo, que se tornou o Comitê de Supervisão (analisado em detalhes no Capítulo 3.3), tendo anunciado os primeiros 20 integrantes do Comitê em maio de 2020, e começado a analisar os primeiros casos em dezembro de 2020. Naquele momento, o Facebook já contava com 15 mil moderadores de conteúdo espalhados no mundo[188].

PUC-Rio - Certificação Digital Nº 2112697/CA

[186] Disponível em: <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>. Acesso em: 20 nov. 2022.
[187] Disponível em: <https://about.fb.com/news/2018/04/comprehensive-community-standards/>. Acesso em: 20 nov. 2022.
[188] Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2020/11/21/moderadores-de-conteudo-do-facebook-enviam-carta-aberta-a-zuckerberg.htm>. Acesso em: 20 nov. 2022.

Capítulo 2 – Os desafios da moderação de conteúdo exercida pelas plataformas de redes sociais

Como visto no Capítulo 1, a atividade de moderação de conteúdo praticada pelas plataformas digitais é necessária para assegurar a liberdade de expressão. Com o tempo, contudo, devido principalmente à falta de transparência e à opacidade das decisões das plataformas, essas empresas passaram a ser criticadas, gerando uma crise de legitimidade de suas decisões.

Há quem defenda que as plataformas interferem de forma exagerada no discurso público, limitando a liberdade de expressão no espaço *online*, e há quem, em sentido oposto, entenda que as plataformas interferem pouco, permitindo a existência de um ambiente digital tóxico, com conteúdo indesejável circulando.

De fato, a atividade de moderação de conteúdo na internet não é simples. Moderar conteúdo é difícil não apenas porque demanda recursos, mas principalmente porque exige que as plataformas tomem decisões difíceis e rápidas sobre aspectos que muitas vezes não são facilmente identificáveis como ilícitos ou ilegais. A dificuldade na interpretação de determinado conteúdo se mostra ainda mais desafiadora ao se considerar que as plataformas atuam em nível global, para bilhões de usuários, e com milhões de conteúdos sendo divulgados a cada dia.

Dessa forma, nesse capítulo serão apresentados os principais desafios enfrentados pelas plataformas de redes sociais na atividade de moderação de conteúdo.

PUC-Rio - Certificação Digital Nº 2112697/CA

2.1 A atuação global das plataformas de redes sociais e o desafio de escala na moderação de conteúdo

Como já visto ao longo desse estudo, as plataformas de redes sociais enfrentam atualmente o desafio de escala, por moderar conteúdo gerado por bilhões de usuários espalhados pelo globo, em contextos culturais, sociais e econômicos distintos[189]. Com efeito, as redes sociais atuam de forma global e têm de lidar com

[189] "Dada a escala em que o Twitter está, uma chance em um milhão acontece 500 vezes por dia. É o mesmo para outras empresas que lidam com esse tipo de escala. Para nós, casos extremos, aquelas situações raras que provavelmente não ocorrerão, são mais como normas. Digamos que 99,999% dos tweets não representem risco para ninguém. Não há nenhuma ameaça envolvida. Depois de retirar esses 99,999%, essa pequena porcentagem de tweets restantes chega a cerca de 150.000 por mês. A escala absoluta do que estamos lidando é um desafio". Del Harvey, vice presidente de Trust

um volume de discurso sem precedentes, o que torna a revisão humana do conteúdo, por exemplo, impossível[190]. Exigir a revisão humana de cada conteúdo postado por usuários demanda tempo, o que não se tem na internet, pois um conteúdo indesejável pode viralizar em questão de segundos.

Como afirma Tarleton Gillespie, plataformas de redes sociais não são apenas grandes. Embora essas plataformas geralmente se apresentem como uma "comunidade", é difícil imaginar uma comunidade quando se tem bilhões de usuários ativos. O autor esclarece que as plataformas gerenciam diferentes comunidades mutáveis, em várias nações, culturas e religiões, muitas vezes com valores e objetivos distintos, sendo que essas comunidades não coexistem independentemente em uma plataforma, elas se sobrepõem e se misturam[191].

As plataformas, usualmente, elaboram suas regras sobre moderação de conteúdo para serem aplicáveis globalmente, e seus termos de uso ou padrões de comunidade costumam ser formulados dessa forma. Regras globais são necessárias por razões de eficiência e completude da experiência "sem fronteiras" de uma rede social[192].

No entanto, o contexto importa quando se fala em liberdade de expressão e na tradução de regras sobre o discurso, que são, como padrão, criadas por poucas pessoas de contextos homogêneos (geralmente concentradas no Vale do Silício), para valer em todo o mundo. Com efeito, não existe uma compreensão "correta" da liberdade de expressão. Algo que é considerado indesejável ou ilícito no Brasil pode não ser em outro país[193]. O Facebook já observou, por exemplo, que "decidir se uma propaganda é política e fazer essa definição funcionar em diferentes jurisdições não

PUC-Rio - Certificação Digital Nº 2112697/CA

and Safety, Twitter, na TED talk, "Protecting Twitter Users (Sometimes from Themselves)," Março, 2014, *In*: GILLESPIE, Tarleton. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 74.

[190] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 81.

[191] GILLESPIE, Tarleton. op. cit., p. 76.

[192] MONIKA, Bickert. "Defining the boundaries of free speech on social media". *In: The free speech century*. Edited by Lee C. Bollinger and Geoffrey R. Stone. 2019, p. 262.

[193] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 10 nov. 2022.

é banal"[194]. Ainda, como observa Monika Bickert, *head of policy management* do Facebook:

> A praticidade de implementar padrões em comunidades tão grandes simplesmente requer uma mão pesada das empresas. Mesmo que os padrões de discurso on-line fossem definidos por uma autoridade, o nível de atenção necessária para implementar qualquer padrão em uma escala tão grande significa que as empresas devem desempenhar um papel primordial na decisão final de remover ou deixar no site qualquer conteúdo[195].

Ademais, casos envolvendo conflitos de liberdade de expressão abrangem a colisão com outros direitos fundamentais. Regular a liberdade de expressão, especialmente em casos de discurso de ódio e nudez, por exemplo, cria conflitos com a liberdade e a igualdade, pois a liberdade de expressão de uma pessoa pode ter impacto em direitos de terceiros. Como não há uma compreensão correta sobre a liberdade de expressão, restrições sobre o discurso devem ser esclarecidas com antecedência, mas devem também serem aplicadas com especial atenção ao contexto em que o discurso foi feito. Por isso mesmo, regulações nacionais importadas para outros contextos causam problemas na governança sobre o discurso[196]. Sobre esse desafio, assim dispõe o relatório Olhando Al Sur:

> Por outro lado, o fato de que as plataformas da Internet sejam responsáveis pela determinação da legalidade ou ilegalidade de uma expressão traz em si um problema que viola os princípios democráticos. Significa a delegação de poderes do Estado às empresas privadas, conferindo-lhes o poder de decidir qual expressão é lícita e qual não é, sem a devida transparência, responsabilidade ou procedimentos efetivos de recurso e/ou reparação[197].

Para enfrentar o desafio de escala, tornou-se necessária a criação de ferramentas automatizadas de inteligência artificial, que são atualmente determinantes na moderação de conteúdo. Tais ferramentas, contudo, são naturalmente passíveis de erros, e não têm a mesma precisão na capacidade de

PUC-Rio - Certificação Digital Nº 2112697/CA

---

194 Cf.: Protecting Election Integrity. Facebook, [s.d]. Disponível em: <https://www.facebook.com/business/m/election-integrity>. Acesso em: 20 nov. 2022.

195 Tradução livre de: "The practicality of implementing standards in communities this large simply requires a heavy hand from the companies. Even if online speech standards were set by an outside authority, the level of attention required to implement any standard at such a large scale means that companies must play a primary role in the ultimate decision to remove or leave on site any given piece of content". MONIKA, Bickert. "Defining the boundaries of free speech on social media". *In: The free speech century*. Edited by Lee C. Bollinger and Geoffrey R. Stone. 2019, p. 257.

196 DOUEK, Evelyn. *Verified Accountability*. Disponível em: <https://www.hoover.org/sites/default/files/research/docs/douek_verified_accountability_aegisnstl1903_webreadypdf.pdf>. Acesso em: 10 nov. 2022.

197 CAMPO, Augustina Del, et al. *Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet*. Abril, 2021. Disponível em: <https://www.alsur.lat/sites/default/files/2021-06/Responsabilidad%20de%20intermediarios%20PT.pdf>. Acesso em: 10 nov. 2022.

interpretar o contexto de uma publicação como teriam revisores humanos, por exemplo.

Tarleton Gillespie afirma que a detecção automatizada não é uma tarefa fácil, pois a ofensa depende tanto da interpretação quanto do contexto, e algoritmos de identificação automática têm dificuldade em discernir conteúdo ou comportamento ofensivo. A identificação automatizada se torna ainda mais complicada quando as plataformas tentam detectar se algo é pornografia ou discurso de ódio, sem conseguir compará-lo a um catálogo de exemplos[198].

Ao errar em identificar o real significado de uma publicação, essas ferramentas de inteligência artificial, que são necessárias, acabam criando mais um desafio de possível equívoco ou até mesmo de viés discriminatório na decisão, podendo afetar o direito à liberdade de expressão dos usuários.

Estudo conduzido pelo InternetLab revelou que ferramentas algorítmicas desenvolvidas por plataformas de redes sociais para aferir a "toxidade" de uma publicação podem não ser capazes de diferenciar, por exemplo, conteúdo de ódio dirigido à comunidade LGBTQIA+ do conteúdo publicado pelos próprios membros dessa comunidade, utilizando-se de linguagem pseudo-ofensiva como forma de protesto[199]:

> Portanto, operando em contextos culturais, políticos e sociais radicalmente diversos e lidando com milhões de usuários, as equipes de elaboração de políticas e diretrizes, os sistemas automatizados e equipes de revisores não têm dado conta de realizar a moderação sem cometer erros e abusos à liberdade de expressão de seus usuários ou reforçar discursos discriminatórios e violentos ilegais[200].

Exemplo brasileiro desse desafio pode ser visto no caso de um usuário que publicou uma foto no Instagram em 2020 com imagens de sintomas de câncer de mama, exibindo mamilos, com título em português indicando apoio ao Outubro Rosa. O *post* foi removido pelo sistema automatizado do Facebook, aplicando as políticas do Padrão da Comunidade de Supervisão sobre Nudez Adulta e Atividade Sexual do Facebook[201], muito embora tais políticas tragam como exceção, de forma

[198] GILLESPIE, Tarleton. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 107.
[199] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 10 nov. 2022.
[200] Ibid.
[201] "Comitê de Supervisão revoga decisão original do Facebook: caso 2020-004-IG-UA", Comitê de Supervisão, Notícias (blog), janeiro de 2021. Disponível em: <oversightboard.com/>. Acesso em: 02 dez. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

expressa, publicações de mamilos para conscientização sobre o câncer de mama. Esse caso foi para o Comitê de Supervisão criado pelo Facebook em 2020 para que reavaliasse a decisão, de forma independente[202]. O Comitê revogou a decisão, enfatizando o fato de a automação criar riscos para liberdade de expressão, como no caso em questão, eis que, até a decisão pela restauração do conteúdo, já havia passado o mês da campanha[203].

Sobre o uso de ferramentas automatizadas para lidar com o problema de conteúdo postado em escala, assim observa o relatório Olhando Al Sur:

> Na moderação de conteúdo em larga escala, exigir que as empresas atuem com celeridade ou "sem atrasos indevidos" as impede de realizar uma análise adequada de cada caso e, ao mesmo tempo, indiretamente as pressiona a implementar mecanismos de automatização que colaborem à detecção rápida de conteúdo ilegal e problemático. É o caso da norma alemã, por exemplo, e da proposta de Regulamento para a Prevenção da Divulgação de Conteúdo Terrorista Online da União Europeia. A automação da moderação tem seus próprios problemas, que vão desde a má interpretação e a falta de contexto do conteúdo, até preconceitos inerentes que aumentam a discriminação contra grupos vulneráveis. Por isso, embora o investimento em tecnologias de moderação de conteúdo e análise de reclamações seja necessário para a moderação de conteúdo em escala, deve-se promover que os algoritmos e atividades envolvidos sejam realizados de forma adequada, de forma a não afetar a expressão dos usuários[204].

No entanto, embora a automação contenha erros, é necessária para lidar com o problema de escala, pois é tecnicamente impossível que todo conteúdo publicado em redes sociais seja revisado por humanos, em um tempo razoável (como se sabe, um conteúdo pode viralizar em questão de segundos), para reduzir o impacto de um conteúdo ofensivo.

De toda forma, o uso dessas ferramentas de inteligência artificial não dispensa, por completo, o trabalho dos revisores humanos. Embora o uso de ferramentas automatizadas seja necessário para lidar com o problema de escala, revisores humanos também permanecem sendo indispensáveis, pois, até o momento, apenas uma pessoa é capaz de interpretar o contexto de um texto, imagem ou vídeo para compreender seu real significado e pertinência conforme as regras

[202] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 22 nov. 2022.

[203] Disponível em: <https://www.oversightboard.com/sr/decision/004/Portuguese>. Acesso em: 20 nov. 2022.

[204] CAMPO, Augustina Del. et al. *Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet*. Abril, 2021. Disponível em: <https://www.alsur.lat/sites/default/files/2021-06/Responsabilidad%20de%20intermediarios%20PT.pdf>. Acesso em: 10 nov. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

estabelecidas. Afinal, "uma postagem com uma imagem de um ataque terrorista pode veicular uma crítica ou uma exaltação, a depender das palavras que a acompanham"[205].

Diante desse desafio de escala, qualquer proposta de regulação normativa ou regulatória sobre a atividade de moderação de conteúdo exercida pelas redes sociais deve levar em consideração a complexidade desse novo contexto tecnológico. Como explica Lawrence Lessig, "a lição mais importante sobre o direito no ciberespaço é a necessidade de se levar em conta o efeito regulatório do código"[206]. Não só os riscos existentes nesse efeito regulatório do código, mas também as necessidades legítimas que se busca enfrentar[207].

2.2 Desafios do baixo *accountability* e da transparência insuficiente na moderação de conteúdo exercida pelas plataformas de redes sociais

PUC-Rio - Certificação Digital Nº 2112697/CA

Além dos desafios de falha de identificação de conteúdo indesejável ou ilegal diante da escala e das decisões automatizadas, as plataformas de redes sociais enfrentam, ainda, desafios relacionados ao baixo *accountability* e à transparência insuficiente na tomada de suas decisões. As plataformas de redes sociais são constantemente criticadas pela opacidade dos algoritmos de moderação de conteúdo, dos sistemas de recomendação que determinam a ordem e o tipo de conteúdo veiculado aos usuários, bem como da aplicação de seus termos de uso[208].

De fato, embora tenha havido movimentos por mais transparência das plataformas nos últimos anos, a aplicação de suas regras internas ainda ocorre majoritariamente de forma opaca. A maior parte das plataformas de redes sociais revelam, por meio de relatórios de transparência, números agregados sobre as decisões tomadas contra vários tipos de conteúdo, mas sem fornecer exemplos

[205] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 82.
[206] LESSIG, Lawrence. *Code*: *version 2.0*, Basic Books, 2006, p. 155.
[207] NITRINI, Rodrigo Vidal. op. cit.
[208] LEERSEEN, Paddy. *The soap box as a black box: regulating transparency in social media recommender systems.* European Journal of Law and Technology, v. 11, n. 2, 2020. Disponível em: <https://www.ejlt.org/index.php/ejlt/article/download/786/1012/3408>. Acesso em: 22 nov. 2022.

concretos ou maiores informações sobre as decisões, tornando difícil avaliar o que esses relatórios de fato significam[209].

Por isso, é quase unanimidade entre estudiosos do tema que as plataformas devem ter obrigações de transparência sobre as suas decisões. Afinal, sem transparência não há como os usuários exercerem outros direitos procedimentais, como o devido processo legal e a isonomia[210].

Assim também vem entendendo a jurisprudência brasileira. O Tribunal de Justiça de São Paulo ("TJSP"), por exemplo, recentemente condenou o Twitter pelo bloqueio de contas da plataforma sem a indicação das "causas concretas que resultaram na aplicação da penalidade e sem viabilizar o exercício do regular contraditório", o que retirou da autora "o direito de conhecer as tais irregularidades no tempo oportuno, vindo a ser surpreendida, desde logo e mediante juízo absolutamente discricionário, com a aplicação da sanção mais severa, de cancelamento/exclusão da conta"[211].

De forma semelhante, no caso da suspensão do canal "Terça Livre TV" pelo Google, na plataforma YouTube, o TJSP entendeu pela improcedência do pedido do autor Allan Lopes dos Santos pela condenação do Google ao restabelecimento da conta suspensa, justamente considerando que na "ocasião, o apelante foi avisado de que eventuais tentativas de burla da suspensão por meio de outros canais poderiam levar à desativação das contas, nos termos de serviço da plataforma", mas seguiu violando os termos de uso da plataforma[212].

A falta de dados divulgados pelas plataformas sobre a atividade de moderação de conteúdo gera impactos tanto na perspectiva do interesse público, quando no direito de defesa dos usuários. Para estes, a falta de informações básicas sobre, por exemplo, os motivos e critérios que levaram à remoção de um conteúdo

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[209] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 22 nov. 2022.

[210] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 242.

[211] BRASIL. TJSP. Apelação Cível nº 1058263-04.2020.8.26.0100. 22ª Câmara de Direito Privado, Rel. Des. Edgar Rosa, j. em 03/03/2021.

No mesmo sentido: BRASIL. TJSP. Apelação Cível 1105667-22.2018.8.26.0100. Rel. Des. Morais Pucci. 35ª Câmara de Direito Privado. j. em 27/07/2020; TJSP. Apelação Cível 1002982-60.2019.8.26.0565. Rel. Des. Silvério da Silva. 8ª Câmara de Direito Privado. j. em 05/07/2020.

[212] BRASIL. TJSP. Apelação Cível nº 1073111-59.2021.8.26.0100. Rel. Des. Marcus Vinicius Rios Gonçalves. 6ª Câmara de Direito Privado. j. em 26/05/2022.

e possibilidade de recursos para contestar uma decisão, tornam praticamente impossível a apropriada defesa por parte do usuário, além de esconder possíveis arbitrariedades e erros nas decisões da plataforma, que podem acabar limitando a liberdade de expressão dos usuários de forma equivocada[213]. David Kaye, Relator Especial das Nações Unidas para a Liberdade de Opinião e Expressão entre 2014 e 2020, esclarece que falhas no cumprimento de deveres de transparência ameaçam a capacidade dos usuários de compreenderem os limites impostos à sua liberdade de expressão e de buscarem os remédios adequados contra decisões que entendam equivocadas[214].

Na perspectiva do interesse público, plataformas privadas influenciam no que é ou não visto e falado mediante filtragem, etiquetagem e exclusão de conteúdo e conta, e, por isso, devem cumprir compromissos mínimos de transparência para informar a sociedade sobre as decisões tomadas[215]. Com efeito, garantir maior transparência aprimora o debate público sobre a aplicação das políticas e as escolhas tomadas pelas plataformas de redes sociais, permitindo um maior controle, inclusive pelo Poder Judiciário, sobre as plataformas e suas decisões. Luna Barroso explica que a garantia da transparência atende a três objetivos fundamentais:

> [E]m primeiro lugar, fornece aos usuários maior compreensão e conhecimento sobre se e em que medida as plataformas atuam para regular discurso, mantê-los seguros no ambiente digital e prevenir danos; em segundo lugar, garantem ao judiciário ou ao órgão regulador designado, e a pesquisadores, maiores informações para compreenderem as ameaças dos serviços digitais, o papel das plataformas na amplificação ou minimização desses riscos, e eventuais ações de mitigação de danos adotadas; em terceiro lugar, servem para garantir que as plataformas terão algum tipo de accountability público sobre suas decisões de moderação de conteúdo e sobre os impactos de seus serviços, promovendo debate qualificado que busque aprimorar as práticas da indústria como um todo[216].

O aperfeiçoamento da transparência, contudo, também envolve complexos desafios. Com efeito, fornecer informações detalhadas sobre as decisões de moderação de conteúdo pode acabar tendo efeito reverso, provendo subsídios a pessoas mal-intencionadas que terão, com eles, capacidade de burlar a forma como

---

[213] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 22 nov. 2022.

[214] HUMAN RIGHTS COMMITTEE. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. 11 mai. 2016. UN Doc A/HRC/32/38. Disponível em: <https://undocs.org.en/A/HRC/32/38>. Acesso em: 22 nov. 2022.

[215] VALENTE, Mariana G. et al. Op. cit.

[216] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 243.

PUC-Rio - Certificação Digital Nº 2112697/CA

as regras são aplicadas, o que é conhecido como *gaming the system*. Se contas que disseminam *spam* souberem como as plataformas detectam tais mensagens não-solicitadas, por exemplo, saberão como burlar essa regra e impedir que as plataformas removam esse conteúdo, lotando os *feeds* dos usuários com conteúdo indesejado ou fraudulento[217].

Ademais, exigências de transparência podem causar impactos sobre outros interesses relevantes, como a garantia da privacidade. Há, por exemplo, a legítima necessidade de se evitar excessiva transparência em casos de discurso de ódio, *bullying* etc. Republicar e dar maior visibilidade para conteúdo de discurso de ódio ou mensagens de *bullying* como parte de um esforço de "transparência" pode acabar fazendo o trabalho do usuário que a plataforma decidiu que não deveria ser ouvido[218].

Além disso, existe a complexidade na definição de quais informações devem ser públicas e como elas devem ser apresentadas, contextualizadas e acessadas. Acreditar que quanto mais informação, melhor, não é correto, pois exigências de transparência excessivas incluem custos de oportunidades e diversos riscos. De fato, tais exigências podem fazer com que as plataformas adotem regras sobre moderação de conteúdo mais simples, na tentativa de reduzir seus custos de classificação e fundamentação das decisões, o que pode potencialmente levar a um excesso ou insuficiência na remoção de conteúdo, viabilizando um ambiente digital tóxico. Ainda, exigências exacerbadas de transparência podem barrar a entrada de pequenas e médias empresas, pelo custo gerado. Por fim, impor exigências de transparência de forma padronizada para todas as plataformas é prejudicial à liberdade de expressão, pois limita as alternativas para usuários insatisfeitos com as regras de uma plataforma[219].

Como a transparência é um importante bem instrumental para se alcançar maior *accountability* às plataformas, os desafios envolvidos na transparência,

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[217] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 22 nov. 2022.

[218] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 22 nov. 2022.

[219] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 246.

incluindo os custos para aplicação de determinadas medidas, devem ser considerados de forma específica pelos reguladores, sem referência a conceitos abstratos como "transparência", apenas[220]. É mais importante um foco sistêmico que revele informações úteis do que um foco em transparência pública em cada caso concreto[221].

2.3 Desafio da legitimidade das decisões tomadas pelas plataformas de redes sociais na moderação de conteúdo

A atividade de moderação de conteúdo praticada pelas plataformas de redes sociais, muito por conta das dificuldades de escala, do baixo *accountability* e da falta de transparência, enfrenta o desafio, ainda, da legitimidade das decisões.

Em 2017, um morador de Cleveland matou a tiros um idoso negro que andava perto de sua casa. Ele disse que decidiu matar o idoso pois havia brigado com sua ex-namorada, e estava estressado. O vídeo do crime, postado no Facebook pelo próprio atirador, ficou no ar por duas horas antes de ser removido pela plataforma[222]. Em 2016, outro cidadão negro, Philando Castile, foi baleado durante uma abordagem policial dentro de seu carro. Sua namorada transmitiu um vídeo ao vivo no Facebook em que ele agonizava até morrer[223]. O vídeo foi rapidamente removido da plataforma, mas foi posteriormente restabelecido sob o aviso prévio de que se tratava de material com cenas fortes. Essa transmissão ao vivo teve repercussão nacional, sob o tema da violência policial a pessoas negras nos Estados Unidos, na esteira do movimento Black Lives Matter[224].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[220] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at. Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 22 nov. 2022.

[221] EDWARDS, Lillian; VEALE, Michael. "*Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*," Duke Law & Technology Review 16, no. 1 (2017), p. 67.

[222] Disponível em: <https://agora.folha.uol.com.br/mundo/2017/04/1876437-assassino-nos-eua-mata-idoso-ao-vivo-no-facebook.shtml>. Acesso em: 22 nov. 2022.

[223] Disponível em: <https://www.washingtonpost.com/news/the-switch/wp/2016/07/07/why-facebook-took-down-the-philando-castile-shooting-video-then-put-it-back-up/>. Acesso em: 22 nov. 2022.

[224] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 114.

Embora os dois vídeos possam ser similares, a empresa eventualmente determinou que as intenções por detrás de cada um lhe conferia significados diferentes. Deixar o segundo vídeo no ar chamava atenção ao racismo institucional da polícia americana, enquanto remover o primeiro "silenciava uma homenagem insana de um assassino a sua ex-namorada"[225].

Em 2016, durante a campanha presidencial do então candidato Donald Trump, o Facebook já discutia internamente se permitia ou proibia postagens do ex-Presidente que propunham a proibição da entrada de muçulmanos nos Estados Unidos, sob o argumento de que "grande parte dos muçulmanos odeia os americanos". Diversas dessas postagens foram denunciadas como inapropriadas por usuários, e parte considerável da equipe de moderação de conteúdo da empresa entendia que elas violavam as regras contra discursos de ódio e pedidos de exclusão de um grupo religioso[226]. Ao final, as publicações de Trump foram permitidas, mas as razões dessa decisão não foram levadas a público, e, assim, nunca foram esclarecidas.

Noutra toada, em 2018, a conta do filho do primeiro-ministro de Israel foi temporariamente suspensa depois que ele publicou que desejava que "todos os muçulmanos saíssem da terra de Israel". A suspensão da conta foi feita com base nas regras de vedação ao discurso de ódio do Facebook[227].

Esses exemplos de decisões editoriais do Facebook demonstram que as grandes plataformas de redes sociais alcançaram a difícil posição de árbitros na aplicação de suas próprias regras na seara de debates públicos, e no político e eleitoral, em especial, definindo o que pode ou não circular nas redes. Ocorre que, como já visto, o processo decisório das plataformas ainda é opaco, o que acaba dando margem para que as decisões sejam percebidas como arbitrárias ou enviesadas. Isso, somado a um histórico de inconsistência nas decisões e de idas e vindas nas formulações de regras, contribuíram para o desafio da legitimidade das decisões tomadas pelas plataformas[228].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[225] KLONICK, Kate. *Inside the team at Facebook that dealt with the Christchurch shooting*, The New Yorker, artigo publicado em 25/04/2019.
[226] NITRINI, Rodrigo Vidal. op. cit., p. 114-115.
[227] Facebook temporarily band Israeli PM's son over post, reportagem publicada pela BBC News, em 17/12/2018. KADRI, Thomas E., KLONICK, Kate. Facebook v. Sullivan: *public figures and newsworthiness in online speech*. Southern California Law Review, v. 93, p. 37-99, 2019, p. 92.
[228] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 119.

Ao mesmo tempo em que as plataformas são cobradas para fazer mais para manter um ambiente digital sadio, são criticadas por interferirem em excesso e limitarem a liberdade de expressão dos usuários[229]. Nos Estados Unidos, por exemplo, há uma crítica do partido republicano arguindo que contas conservadoras estão sendo suspensas pelas plataformas, que seriam controladas por um viés de esquerda. Do outro lado, no partido democrata, existem críticas às plataformas no sentido de que elas não estão sendo ativas para remover conteúdo de discurso de ódio, tóxico e que causa danos aos usuários. Situação semelhante foi vista no Brasil, nas eleições presidenciais de 2022. De um lado, bolsonaristas alegam que as plataformas atuam em excesso, removendo conteúdo de maneira injustificada e censurando o discurso[230]. Do outro, afirma-se que as plataformas não atuam de forma satisfatória para remover conteúdo de desinformação. Ao tratar do tema das *fake news* e eleições, Carlos Affonso Souza e Chiara Spadaccini de Teffé apontam:

> O papel das plataformas no combate da desordem informacional não é isento de controvérsias. Ao atuar para eliminar contas inautênticas e que, de modo artificial procurariam amplificar o alcance de um conteúdo, o Facebook foi alvo de questionamentos e legações de que estaria [limitando] a expressão dos usuários afetados, especialmente de um espectro político de direita[231].

As críticas, como se vê, não batem. De um lado se diz que as plataformas removem demais, e do outro, de menos. Diante da crise de legitimidade das decisões das plataformas, o próprio Mark Zuckerberg tem defendido, desde 2019, um maior nível de regulamentação sobre as plataformas de internet, em especial em quatro áreas: (i) eleições "com padrões comuns para verificar quem é um ator político", e abordar "importantes questões sobre como campanhas políticas usam dados e microdirecionamento", (ii) privacidade, (iii) proteção de dados, e (iv) moderação de conteúdos considerados problemáticos pelas redes sociais[232].

Fato é que a moderação de conteúdo é necessária, e não pode mais ficar nas sombras das grandes plataformas. Deve-se pensar em um modelo de regulação

---

[229] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 22 nov. 2022.

[230] Disponível em: <https://www1.folha.uol.com.br/poder/2022/10/bolsonaristas-gritam-xandao-ditador-e-pedem-fim-da-censura-em-ato-na-paulista.shtml>. Acesso em: 22 nov. 2022.

[231] SOUZA, Carlos Affonso Souza; TEFFÉ, Chiara Spadaccini de. Responsabilidade dos provedores por conteúdos de terceiros na internet. *In:* ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação.* 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021, p. 310.

[232] ZUCKERBERG, Mark. "*The internet needs new ruled.* Let's start in these four areas", artigo publicado pelo The Washington Post, em 30/03/2019.

PUC-Rio - Certificação Digital Nº 2112697/CA

capaz de desenvolver mecanismos de legitimação das decisões relacionadas ao exercício da liberdade de expressão, promovendo credibilidade e transparência aos processos decisórios sobre moderação de conteúdo.

2.4 Desafios advindos da regulação estatal sobre a moderação de conteúdo exercida pelas plataformas de redes sociais

Diante dos desafios apresentados acima, é possível concluir que governos não devem apenas aplicar regras locais para regular o discurso *online*, que se faz presente de forma global. A regulação estatal pode fornecer limites permissíveis em um sentido amplo, de acordo com as restrições de suas constituições. No entanto, a regulação estatal sobre moderação de conteúdo não fornece uma resposta acurada na maior parte dos casos difíceis que as plataformas têm para decidir, pelas razões elencadas a seguir[233].

Em primeiro lugar, a regulação estatal não conseguirá gerenciar a velocidade e a escala em que as plataformas têm que decidir difíceis questões. A regulação sobre o discurso *online* demanda uma atuação pesada das plataformas para traduzir regras nos casos concretos em um ambiente dinâmico. Governos podem definir *standards* amplos para serem observados, mas devem parar por aí. Normas *online* de discurso, memes e linguagem mudam a cada dia, ou mesmo hora. Os processos governamentais dificilmente serão ágeis ou responsivos o suficiente para gerenciar a rápida evolução das disputas de fala *online*[234].

Em segundo lugar, ainda que fosse possível na prática que regulações estatais gerenciem os desafios de escala e do volume de conteúdo das redes sociais, a noção de governos terem esse envolvimento na regulação do discurso está em tensão com os propósitos democráticos da liberdade de expressão[235], podendo configurar censura estatal. Sobre esse ponto, David Kaye, então Relator Especial das Nações Unidas para a Liberdade de Expressão, já destacou que:

PUC-Rio - Certificação Digital Nº 2112697/CA

[233] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at. Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 22 nov. 2022.
[234] Ibid.
[235] Ibid.

> [L]eis restritivas formuladas com termos vagos como 'extremismo', blasfêmia, difamação, discurso 'ofensivo', 'notícias falsas' e 'propaganda' costumam servir como pretextos para exigências de que as empresas suprimam discurso legítimo. Cada vez mais, os Estados têm como alvo específico conteúdo publicado em plataformas digitais[236].

Em terceiro lugar, há uma tendência de as plataformas removerem em excesso se tiverem o risco de serem responsabilizadas pela existência de determinado conteúdo tido por ilícito pelo governo. Com efeito, um maior grau de responsabilização civil de plataformas por conteúdo gerado por terceiro cria um incentivo para a derrubada de conteúdo, o que é problemático do ponto de vista da liberdade de expressão[237], podendo constituir censura colateral. Como explica o relatório Olhando Al Sur, ameaças de responsabilidade, incluindo multas significativas e até mesmo prisão, como no caso da legislação australiana, "somadas à pressão para resolver em prazos extremamente curtos, geram um incentivo para a retirada excessiva de conteúdo, conhecido como censura 'privada'", e, diante disso, "teme-se que as plataformas eliminem conteúdos supostamente ou manifestamente ilegais e, em muitos casos, totalmente legais, violando a proteção ao direito à liberdade de expressão reconhecida em instrumentos internacionais"[238].

Luna Barroso esclarece que esse tipo de censura aparece quando o Estado regula ou obriga uma plataforma de rede social a restringir o discurso de um terceiro que não integra a estrutura da plataforma. Como a empresa não tem relação com esse terceiro ou proveito específico na defesa do conteúdo, não terá qualquer interesse em litigar contra o Estado pela manutenção do conteúdo, ainda em casos que se entenda tratar de conteúdo perfeitamente lícito. Assim, para evitar sanções ou ameaças regulatórias, as plataformas tenderão a remover o conteúdo sem maiores questionamentos e sem oportunizar ao autor defesa pela licitude do conteúdo. A remoção de conteúdo perfeitamente lícito pode criar espécie de efeito

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[236] HUMAN RIGHTS COMMITTEE. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. 11 mai. 2016. UN Doc A/HRC/32/38. Disponível em: <https://undocs.org.en/A/HRC/32/38>. Acesso em: 22 nov. 2022.

[237] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 175.

[238] CAMPO, Augustina Del, SCHATZKY, Morena, HERNÁNDEZ, Laura, LARA, J. Carlos. *Olhando Al Sur*. Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet. Abril, 2021. Disponível em: <https://www.alsur.lat/sites/default/files/2021-06/Responsabilidad%20de%20intermediarios%20PT.pdf>. Acesso em: 22 nov. 2022.

silenciador (*chilling effect*) na versão digital[239]: "não são os oradores que deixarão de falar por medo, mas as plataformas que restringirão e filtrarão conteúdo em excesso para evitar a aplicação de sanções"[240].

Em quarto lugar, deveres regulatórios muito exigentes também podem acabar criando ainda mais concentração ao estabelecer, por exemplo, exigências que apenas as empresas de capital bilionário serão capazes de atender. "Isso teria o efeito contrário ao buscado, dado que uma parte importante das preocupações com moderação de conteúdo está justamente no acúmulo de poder num pequeno número de empresas"[241].

Por fim, regulações estatais definindo os limites da liberdade de expressão podem colocar em perigo o projeto de plataformas globais, que atuam em escala. O Facebook, por exemplo, para permitir e preservar o diálogo entre seus usuários ao redor do mundo, precisa garantir que todos tenham acesso ao mesmo conteúdo. Por isso, regulações estatais que colocam "limites à liberdade de expressão em plataformas digitais podem representar limitações a essa troca global. Quando um país define o que pode ou não ser dito no Facebook, a comunicação entre usuários em jurisdições diferentes pode vir a ser prejudicada"[242].

Por isso mesmo, como se defenderá no Capítulo 3, a regulação estatal sobre moderação de conteúdo deve focar em legitimar o *processo* pelo qual as decisões das plataformas são tomadas, e não focar no aspecto *substancial* dessas decisões. Ademais, regulações estatais punitivas criam incentivos para derrubada excessiva de conteúdo. Por isso, considera-se que a concorrência de um sistema judicial com um autônomo das plataformas evitaria esses incentivos pela supressão de conteúdo[243].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[239] BALKIN, Jack M. *Free Speech is a Triangle*. Columbia Law Review, v. 118, n. 07, p. 2011/2056, 2018. Disponível em: <https://policyreview.info/pdf/policyreview-2019-2-1407.pdf>. Acesso em 12 out. 2022.

[240] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 100.

[241] VALENTE, Mariana G. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo*", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 22 nov. 2022.

[242] ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. *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/>. Acesso em: 07 jul. 2021.

[243] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 175.

2.5 Exemplos de tentativas de regulação estatal da moderação de conteúdo exercida pelas plataformas de redes sociais

2.5.1 Alemanha

Em junho de 2017, o Congresso Alemão aprovou a NetzDG ("*Network Enforcement Act"*), lei aplicável às plataformas de redes sociais e outros provedores, desde que tenham pelo menos 2 milhões de usuários registrados na Alemanha. A lei traz obrigações de transparência e de remoção de conteúdo ilegal, considerado como todo conteúdo que viole dispositivos do código penal alemão.

A lei obriga as plataformas a removerem conteúdo "manifestamente ou obviamente ilegal" dentro de do prazo de 24 horas após serem notificadas, e toma como parâmetro cerca de vinte seções do código penal alemão, que incluem um amplo rol de delitos.

A lei alemã faz uma distinção entre conteúdo "ilegal" e "manifestamente ilegal". Enquanto o prazo para a retirada do segundo é de 24 horas, para o primeiro é de sete dias. Caso deixem de remover o conteúdo no prazo legal, as plataformas poderão ser multadas em até 50 milhões de euros. Além disso, a lei exige que empresas como a Meta criem um procedimento próprio para receber e processar esse tipo de notificação, tomando uma decisão dentro do prazo estipulado[244], e devendo informar às pessoas envolvidas as razões de cada decisão. Na prática, a lei abandonou a ideia da imunidade legal (*safe harbour*) da Seção 230, do CDA, dos Estados Unidos, aos intermediários digitais[245].

Ao determinar essa obrigação de remoção para as plataformas, a lei parece ter criado (o que será melhor avaliado com o tempo) um incentivo para que elas pequem pelo excesso, removendo todo o conteúdo que possa, potencialmente, se enquadrar como "ilegal" ou "manifestamente ilegal", para cumprir o prazo indicado

[244] ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. *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/>. Acesso em: 07 jul. 2021.

[245] Germany: the act to improve enforcement of the law in social networks, legal analysis pela Article 19, publicada em agosto de 2017.

PUC-Rio - Certificação Digital Nº 2112697/CA

e evitar multa[246]. Ou seja, um dos efeitos colaterais da lei alemã é uma maior limitação à liberdade de expressão dos usuários de redes sociais[247]. Isso parece fazer sentido no contexto alemão, em que, por razões históricas envolvendo o nazismo e o holocausto, a liberdade de expressão não é um direito com posição privilegiada no ordenamento, diferentemente do que ocorre na atual interpretação desse direito pelo STF no Brasil, por exemplo, como visto no Capítulo 1.2.

Além disso, entende-se que a lei alemã acabou por terceirizar às plataformas a tarefa de interpretar o conteúdo postado pelos usuários à luz da legislação alemã, para definir se se trata, ou não, de conteúdo ilegal. Ao assim proceder, a legislação acabou delegando para empresas privadas o poder de interpretar a legislação alemã e decidir se algo é ou não ilegal[248], o que afasta a possibilidade de as plataformas formularem seus próprios juízos de moderação de conteúdo que sejam desprovidos de uma vinculação direta com esses juízos de legalidade, a partir de uma liberdade editorial que responde a variados incentivos sociais.

A forma de moderação de conteúdo exercida pelas redes sociais hoje demonstra que essa imposição feita pela legislação alemã não é necessária, pois a governança privada sobre o discurso pode responder de forma mais eficaz a incentivos legais de natureza não punitiva[249]. A perspectiva de uma governança privada de discursos impede mecanismos de censura colateral por parte dos Estados, e "permite abordar também problemas graves que não guardam correlação

PUC-Rio - Certificação Digital Nº 2112697/CA

[246] Vale destacar, contudo, que alguns estudos recentes sugerem que, na prática, esse temor foi mitigado por dois elementos: (i) o fato de as sanções só serem aplicadas em caso de violações sistêmicas; e (ii) o fato de os Tribunais alemães conhecerem ações que pedem a restituição de conteúdo removido. ZURTH, Patrick. *The German NetzDG as role model or cautionary tale?* Implications for the debate on social media Liability. 31 Fordham Intell, Prop, Media & Ent. L.j. 1084, p. 1130, 2021. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1782&context=iplj>. Acesso em: 22 nov. 2022.

[247] ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. *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/>. Acesso em: 07 jul. 2021.

[248] Em sentido contrário, Martin Eifert defende que: "Sob o pano de fundo do efeito geral, a crítica indignada à NetzDG apresenta-se como uma reação exagerada. A NetzDG deve e pode ser discutida em sua configuração concreta de maneira segura e séria. Entretanto, o desenvolvimento da jurisprudência já tinha demonstrado que o grande desafio atual recai fundamentalmente sobre a configuração adequada de uma responsabilidade fundamentada em regras gerais". EIFERT, Martin. A Lei Alemã para a Melhoria da Aplicação da Lei nas Redes Sociais (NetzDG) e a Regulação da Plataforma. *In*: ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação*. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021, p. 187.

[249] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 179.

com parâmetros de ilegalidades, como é o caso de conteúdos com informações falsas que impliquem riscos sociais relevantes, incluindo campanhas de desinformação"[250].

Ricardo Villas Bôas Cueva critica a NetzDG por tratar apenas dos sintomas do problema, por meio da remoção de conteúdo, mas não ir à sua raiz, "pois não obriga as redes sociais a participar ativamente da persecução penal do criminoso, que seria a única maneira de combater de modo duradouro os discursos de ódio e as notícias falsas"[251].

Em todo caso, a NetzDG trouxe diversos dispositivos que podem ser considerados positivos, por fomentarem a transparência e a prestação de contas por parte das plataformas digitais. É o que entende Rodrigo Nitrini[252]:

> É nesse viés que se encontra uma forte virtude da lei: as obrigações impostas às empresas para criarem canais de reclamações por usuários, formularem análises que levem a conclusões transparentes e publicarem relatórios periódicos que forneçam dados sobre as práticas de moderação de conteúdo são medidas cruciais para uma governança de discursos que permita um nível apropriado de controle social, construção de um devido processo digital e fornecimento de razões para a restrição de direitos fundamentais[253].

A lei também determina que plataformas que recebam mais de 100 milhões de notificações por ano devem publicar relatórios semestrais que detalhem a implementação dos mecanismos previstos pela lei, os critérios usados para avaliar os conteúdos reportados, bem como o número total de reclamações e a quantidade de conteúdos bloqueados, entre outros[254]. Em 2019, por exemplo, o Facebook foi multado em 2 milhões de euros pela autoridade competente, pela publicação de um

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[250] Ibid.

[251] CUEVA, Ricardo Villas Bôas. Alternativas para a remoção de fake news das redes sociais. *In*: ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação*. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021, p. 298.

[252] Noutra toada, Martin Eifert, em defesa da NetzDG, pondera que: "De fato, ela enfrenta algumas preocupações de competência da legislação alemã, mas ela pode também ser um grande impulso motivador para um desenvolvimento europeu. No que toca ao direito constitucional, a lei é muito menos problemática do que a sua lista de acusações deixa aparentar. O perigo de um *overblocking* é discutível. Contudo, esse perito não obstaculiza a postura geral da lei, uma vez que se combate ativamente esse perigo por meio de uma difusão de direitos expressos do usuário e por meio de uma ativação da esfera pública. A NetzDG não deve ser, por isso, desmerecida como um desajuste aberrante resultante de uma política acionista. Ela marca muito mais o início de uma configuração política fortalecida do mundo virtual". EIFERT, Martin. A Lei Alemã para a Melhoria da Aplicação da Lei nas Redes Sociais (NetzDG) e a Regulação da Plataforma. *In*: ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação*. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021, p. 208.

[253] NITRINI, Rodrigo Vidal. Op. cit., p. 178.

[254] Ibid, p. 176.

relatório de transparência incompleto e por deixar a ferramenta de denúncia de publicações de forma "escondida" na plataforma[255].

Tais relatórios, contudo, têm se mostrado insuficientes na prática para análise dos impactos da lei, por conta da falta de padronização entre eles nas diferentes plataformas, e pela falta de acesso ao conteúdo removido, impedindo uma análise do mérito das decisões das empresas. Luna Barroso esclarece que "a mera publicação de índices de remoção pode gerar a falsa impressão de que números mais elevados representam a eficácia da Lei – porém, esse não é necessariamente o caso, já que discurso lícito pode estar sendo removido no bolo"[256].

A NetzDG é considerada um importante símbolo de reação de países europeus contra a dominância de empresas americanas sobre a internet[257]. Embora seus efeitos práticos só possam ser avaliados com o tempo de sua vigência, uma avaliação inicial da lei já aponta que ela serviu, principalmente, para que as plataformas fossem mais diligentes nas aplicações de seus termos de uso, motivadas pelos imperativos de transparência sobre os modos como lidam com as reclamações recebidas[258].

PUC-Rio - Certificação Digital Nº 2112697/CA

2.5.2 França e Reino Unido

Diferente da legislação alemã, as soluções apresentadas pela França e pelo Reino Unido visaram, ao menos em um primeiro momento, estabelecer obrigações procedimentais para as plataformas de redes sociais, se preocupando menos com a ilegalidade de determinado conteúdo.

---

[255] Cf.: Germany fines Facebook for under-reporting complaints. Reuters, 2 jul. 2019. Disponível em: <https://www.reuters.com/article/us-facebook-germany-fine-idUSKCN1TX1IC>. Acesso em: 22 nov. 2022.

[256] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 199.

[257] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 177 .

[258] TWOREK, heidi, LEERSEN, Paddy. *An analysis of Germany's NetzDG Law*. Artigo publicado pelo Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression - Institute for Information Law (Universiteit Van Amesterdam), em abril de 2019.

A França, em 2019, publicou um relatório[259] trazendo contornos sobre um possível marco regulatório para as plataformas de redes sociais no país. Segundo o relatório, seu objetivo é estabelecer um equilíbrio entre a abordagem punitiva (como é a NetzDG), e a preventiva, na busca de tornar o processo de moderação de conteúdo pelas redes sociais mais transparente e *accountable*[260].

De acordo com o documento, a assimetria de informação entre órgãos governamentais e as plataformas de redes sociais justifica a intervenção do poder público na atividade de moderação de conteúdo. De acordo com as autoridades francesas, plataformas como o Facebook desenvolveram sistemas de autorregulação pouco transparentes, na tentativa de evitar regulações estatais. Dessa forma, é preciso adotar medidas governamentais para que o processo de moderação de conteúdo seja informado também pelo interesse público, e não apenas pelo interesse privado das plataformas[261].

O relatório francês reconhece que medidas punitivas podem levar a uma maior censura na internet, terceirizando para empresas privadas a responsabilidade pela interpretação de lei nacional (como feito pela NetzDG). Por isso, o relatório focou mais na proposta de criação de um marco legal com objetivo de estabelecer obrigações de transparência e de "defender a integridade dos usuários", e menos em uma proposta punitiva.

Para alcançar esse objetivo, o documento propõe a criação de um órgão independente, formado por representantes do governo, responsável por monitorar a implementação dessas duas obrigações. Entende-se, assim, que a proposta francesa se concentra em uma perspectiva procedimental, "ou seja, muda o processo de moderação de conteúdo em si e aposta na combinação de autorregulação das empresas com regras mínimas estabelecidas pelo Estado para preservar o interesse público"[262].

Em maio de 2020, contudo, a França aprovou a Lei Avia, que, em linha com a legislação alemã, obriga as plataformas de redes sociais a removerem, em menos

PUC-Rio - Certificação Digital Nº 2112697/CA

[259] Disponível em: <https://thecre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf>. Acesso em: 22 nov. 2022.
[260] ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. *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/>. Acesso em: 07 jul. 2021.
[261] Ibid.
[262] Ibid.

de 24 horas, conteúdo manifestamente ilegal e conteúdo que incite o ódio e a violência, e, em menos de 1 hora, propaganda terrorista. Pouco tempo depois, esta obrigação (dentre outras disposições), foi declarada inconstitucional pelo Conselho Constitucional Francês[263], por atentar contra a liberdade de expressão ao não cumprir os requisitos de necessidade e proporcionalidade ao objetivo perseguido[264].

O Reino Unido, em 2019, publicou o *Online Harms White Paper*[265], propondo um novo sistema de regulação das redes sociais, com a criação de um órgão regulador público e independente. De acordo com o documento, tal órgão será responsável pela implementação de padrões para garantir a segurança dos usuários nas redes sociais, ao mesmo tempo em que deverá se preocupar em proteger a liberdade de expressão no ambiente digital[266].

Para tanto, o relatório prevê a criação de um "dever de cuidado" por parte das plataformas, e na promoção de uma "cultura de transparência, confiança e prestação de contas". Dentre outras atividades, o órgão regulador produzirá "códigos de boas práticas" para as redes sociais, monitorará a implementação do "dever de cuidado", preparará relatórios sobre o processo de moderação de conteúdo e, por fim, promoverá campanhas de educação e conscientização do público sobre os desafios impostos pela liberdade de expressão *online*[267].

Tal relatório se transformou no Projeto de Lei *Online Safety Bill*, com a proposta de criar um novo marco regulatório para proteger a segurança dos cidadãos britânicos na internet e combater diferentes categorias de *online harms* – como conteúdo de cunho terrorista, campanhas de desinformação, etc[268].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[263] Disponível em: <https://pt.wikinews.org/wiki/Fran%C3%A7a:_Conselho_Constitucional_censura_a_lei_Avia>. Acesso em: 10 out. 2022.

[264] CAMPO, Augustina Del, et. al. *Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet*. Abril, 2021. Disponível em: <https://www.alsur.lat/sites/default/files/2021-06/Responsabilidad%20de%20intermediarios%20PT.pdf>. Acesso em: 10 nov. 2022

[265] Disponível em: <https://www.gov.uk/government/consultations/online-harms-white-paper>. Acesso em: 10 out. 2022.

[266] ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. *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/>. Acesso em: 07 jul. 2021.

[267] Ibid.

[268] Disponível em: <https://itsrio.org/wp-content/uploads/2021/07/Protecao-de-Dados-e-Transparencia-em-Moderacao-de-Conte%C3%BAdo.pdf>. Acesso em: 10 out. 2022.

Tal Projeto mantém o mencionado dever de cuidado imposto às plataformas para remover conteúdo danoso ou ilegal, sob fiscalização do órgão de comunicação britânico (*Office of Communications* - "OFCOM")[269]. O OFCOM foi criado em 2003, e é responsável por regular diferentes setores de comunicações e telecomunicações, como serviços postais, rádio e televisão.

O Projeto também impõe obrigações de transparência às plataformas, atribuindo ao OFCOM o poder de definir as informações a serem disponibilizadas nos "relatórios de transparência". Nesses, as empresas que estejam dentro do escopo da nova regulamentação devem prestar contas sobre os tipos de *online harms* que estão enfrentando e quais foram as medidas implementadas para combatê-los[270].

O Projeto, contudo, vem sendo criticado por apresentar diversas exceções às suas regras, incluindo em casos de jornalismo, interesse público, e liberdade de expressão. Com efeito, o Projeto recomenda a remoção de conteúdo que seja "legal, mas ofensivo", o que pode levar ao efeito silenciador do discurso na internet[271].

PUC-Rio - Certificação Digital Nº 2112697/CA

2.5.3 Europa

O *Digital Services Act* ("DSA") foi apresentado pela Comissão Europeia em dezembro de 2020. Trata-se de uma proposta de regulação do Parlamento Europeu e do Conselho da Europa para criar um mercado único para serviços digitais dentro dos limites da União Europeia[272]. O DSA traz uma nova proposta de regulamentação sobre as obrigações e responsabilidades dos intermediários na internet, e traz como objetivo as melhores condições para a prestação de serviços digitais inovadores no mercado interno, contribuir com a segurança *online* e a proteção de direitos fundamentais, e consolidar uma estrutura de governança robusta e durável para a efetiva supervisão dos provedores de serviços intermediários[273].

---

[269] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 199.
[270] Ibid.
[271] Disponível em: <https://www.bbc.com/news/technology-59638569>. Acesso em: 10 out. 2022.
[272] Disponível em: <https://itsrio.org/wp-content/uploads/2021/07/Protecao-de-Dados-e-Transparencia-em-Moderacao-de-Conte%C3%BAdo.pdf>. Acesso em: 10 out. 2022.
[273] Comissão Europeia. Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC.

O texto do DSA foi aprovado para discussão pelo Parlamento Europeu em janeiro de 2022, com emendas ao texto original, e deverá agora ser negociado entre o Parlamento e o Conselho Europeu[274]. Em sua versão original elaborada pela Comissão Europeia, o projeto prevê condições para aplicação da regra de isenção de responsabilidade civil dos intermediários por conteúdo gerado por terceiros, impondo às plataformas obrigações abrangentes de diligência e transparência. O DSA busca uniformizar o procedimento de notificação e retirada previsto na Diretiva de Comércio Eletrônico ("DCE"), pois esta atribuía aos Estados ampla liberdade para regular como isso funciona na prática.

O DSA se aplica em casos de "conteúdo ilegal", definido no Regulamento de forma ampla, como "quaisquer informações que, por si só ou por referência a uma atividade, incluindo a venda produtos ou a prestação de serviços, não estejam em conformidade com o direito da União Europeia ou de um Estado-Membro, independente do objeto ou da natureza específica desse direito". Ou seja, a proposta inclui conteúdo como discurso de ódio, terrorismo, pornografia infantil, compartilhamento ilegal ou não consentido de imagens de conteúdo privado, venda ou comercialização de produtos falsos ou ilegais, divulgação não autorizada de conteúdo protegido por direito autoral etc., sendo tudo considerado ilegal nos termos das leis europeias.

A proposta não inclui, contudo, desinformação, pois não existe lei específica na Europa reconhecendo a ilegalidade desse tipo de conteúdo. Em todo caso, as regras do DSA sobre publicidade, transparência de algoritmos e mitigação de riscos, acabam trazendo responsabilidade para as plataformas também em relação a esse tipo de conteúdo[275].

O DSA mantém o regime de responsabilidade dos intermediários previsto no DCE, no sentido de que estes não serão responsabilizados por danos gerados por terceiros em suas aplicações, salvo se eles falharem em remover ou bloquear o acesso às informações ao terem conhecimento do conteúdo ilegal[276]. Tal

PUC-Rio - Certificação Digital Nº 2112697/CA

Bruxelas, 2020, p. 2. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825>. Acesso em: 10 out. 2022.

[274] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 182.

[275] Ibid, p. 184.

[276] "Artigo 14.o Armazenagem em servidor

1. Em caso de prestação de um serviço da sociedade da informação que consista no armazenamento de informações prestadas por um destinatário do serviço, os Estados-Membros velarão por que a

conhecimento pode se dar por notificações apresentadas por cidadãos ou entidades, desde que suficientemente precisas e fundamentadas, ou por investigação proativa das próprias plataformas.

O DSA estabelece obrigações procedimentais às plataformas, de diligência (*due diligence*), com o fim de garantir um ambiente digital "transparente e seguro". Uma das obrigações nesse sentido é que as plataformas adotem termos e condições de uso com informações claras, não ambíguas, e públicas sobre qualquer procedimento, política, e ferramenta utilizada na atividade de moderação de conteúdo. Também prevê que as plataformas atuem de forma diligente, objetiva e proporcional ao aplicar as restrições de seus termos de uso[277].

O DSA prevê, ainda, a obrigação às plataformas de garantir maior transparência, exigindo que as grandes plataformas disponibilizem relatório anual sobre a moderação de conteúdo realizada, incluindo informações como[278]: (i) número de ordens recebidas das autoridades categorizadas de acordo com o tipo de conteúdo ilegal, e o tempo médio para cumprimento das ordens; (ii) número de notificações apresentadas pelos usuários, de acordo com o tipo de conteúdo ilegal, as medidas tomadas em relação às notificações recebidas, o fundamento para tais medidas e o tempo necessário para sua aplicação; (iii) a moderação de conteúdo exercida por iniciativa das próprias plataformas, com os fundamentos para justificar as medidas tomadas; e (iv) número de reclamações recebidas contra as decisões de moderação de conteúdo[279].

Percebe-se que o Regulamento aposta em princípios como *accountability* e transparência para melhor possibilitar o monitoramento desses serviços pelas autoridades públicas da União Europeia. Para garantir a implementação desse

PUC-Rio - Certificação Digital Nº 2112697/CA

responsabilidade do prestador do serviço não possa ser invocada no que respeita à informação armazenada a pedido de um destinatário do serviço, desde que:
a) O prestador não tenha conhecimento efectivo da actividade ou informação ilegal e, no que se refere a uma acção de indemnização por perdas e danos, não tenha conhecimento de factos ou de circunstâncias que evidenciam a actividade ou informação ilegal, ou
b) O prestador, a partir do momento em que tenha conhecimento da ilicitude, actue com diligência no sentido de retirar ou impossibilitar o acesso às informações". Disponível em: <https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=celex%3A32000L0031>. Acesso em: 8 Jan. 2022.

[277] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 185.

[278] Comissão Europeia. Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Bruxelas, 2020, p. 2. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825>. Acesso em: 10 out. 2022.

[279] BARROSO, Luna van Brussel. Op. Cit., p. 186.

arranjo regulatório, o DSA prevê a criação de um *Digital Services Coordinator* (Coordenador de Serviços Digitais) para cada Estado-membro, e um *European Board for Digital Services* (Conselho Europeu de Serviços Digitais) que dará suporte para atuação dos coordenadores nacionais[280].

O Regulamento prevê multas de até 6% do faturamento das plataformas no ano fiscal anterior por violação de seus dispositivos, e de até 1% no caso de fornecimento de informações incorretas ou incompletas.

Após discussões no Parlamento Europeu, foram acrescentadas algumas emendas ao texto apresentado pela Comissão Europeia. Uma delas prevê que usuários devem poder tomar suas decisões de forma livre, autônoma e informada ao utilizar os serviços das plataformas, impossibilitando que as empresas utilizem meios para impedir ou dificultar essa tomada de decisão. Em relação às disposições sobre transparência, o Parlamento acrescentou obrigação para que as plataformas incluam como dado obrigatório nos relatórios o número de anúncios removidos, rotulados ou desabilitados, e a justificativa para essas decisões. Também dispõe que Estados-Membros não devem impor obrigações adicionais de transparência, além das já previstas[281].

PUC-Rio - Certificação Digital Nº 2112697/CA

2.5.4 Brasil (MP nº 1.068/21, e PL nº 2.630/2020)

Como visto no Capítulo 1.3.2, no Brasil, o Marco Civil da Internet regula a responsabilidade civil de provedores de aplicação por danos causados por conteúdo gerado por terceiros nas plataformas, e não traz nenhuma proibição para que as empresas, por atuação própria, moderem conteúdo que viole as políticas previamente aceitas pelos usuários.

A possibilidade de as plataformas moderarem conteúdo, contudo, foi recentemente alvo de controvérsia, diante da minuta de Decreto Presidencial do então presente Jair Bolsonaro que pretendia proibir a moderação de conteúdo exercida pelas plataformas fora das hipóteses taxativamente previstas. Os artigos 2-B e 2-C previam que provedores de aplicações de internet não poderiam, sem ordem

[280] Disponível em: <https://itsrio.org/wp-content/uploads/2021/07/Protecao-de-Dados-e-Transparencia-em-Moderacao-de-Conte%C3%BAdo.pdf>. Acesso em: 10 out. 2022.
[281] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 187.

judicial: (i) excluir, cancelar ou suspender total ou parcialmente os serviços e as funcionalidades das contas mantidas pelos usuários em seus aplicativos, fora das hipóteses previstas em lei; e (ii) excluir, suspender ou limitar a divulgação de conteúdo gerado por usuário em seus aplicativos sem ordem judicial, fora das hipóteses previstas em lei. Destaca-se que as hipóteses permitidas de moderação de conteúdo sem ordem judicial não incluíam conteúdo de desinformação ou discurso de ódio[282].

O Decreto foi duramente criticado por especialistas[283] por violar o princípio da legalidade, já que o Marco Civil da Internet não proíbe a moderação de conteúdo pelas redes sociais. Ademais, o Decreto foi em sentido contrário das iniciativas mundiais de regulação da moderação de conteúdo, que têm dado especial atenção ao combate à desinformação e ao discurso de ódio.

Destaca-se que o Decreto sequer era necessário, eis que existem diversos casos em que o Judiciário determinou o restabelecimento de conteúdo removido com base nos termos de uso da plataforma. Por isso mesmo, Carlos Affonso Souza explica que, embora as plataformas tenham a "primeira palavra" sobre qual conteúdo fica ou sai de uma plataforma, o Judiciário é quem tem a última[284].

Diante da mobilização social contra o Decreto, sua minuta foi substituída pela Medida Provisória ("MP") nº 1.068/2021. A MP, em síntese: (i) criava alguns requisitos procedimentais para remoção de conteúdo ou exclusão de conta, como a necessidade de garantia de informações claras sobre as políticas e procedimentos de moderação de conteúdo adotados pela plataforma, e garantia de contraditório e ampla defesa; (ii) vedava a possibilidade de as plataformas adotarem critérios de moderação de conteúdo que implicassem "em censura de ordem política, ideológica, científica, artística ou religiosa"; e (iii) vedava a exclusão de contas e perfis, exceto em casos de "justa causa"[285]. As hipóteses de justa causa trazidas pela MP incluíam: (i) nudez ou representações explícitas ou implícitas de atos sexuais; (ii) prática, apoio, promoção ou incitação de crimes contra a vida, pedofilia,

[282] Ibid., p. 151.
[283] Nesse sentido: AFFONSO SOUZA, Carlos. *Decreto de Bolsonaro inverte lógica ao impedir moderação de contas e criar indez do que pode ser removido na internet*. Folha de São Paulo, 20 mai. 2021. Disponível em: <https://www1.folha.uol.com.br/poder/2021/05/decreto-de-bolsonaro-inverte-logica-ao-impedir-moderacao-de-contas-e-criar-index-do-que-pode-ser-removido-na-internet.shtml>. Acesso em: 10 out. 2022.
[284] Ibid.
[285] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 152.

PUC-Rio - Certificação Digital Nº 2112697/CA

terrorismo, tráfico ou quaisquer outras infrações penais sujeitas à ação penal pública incondicionada; (iii) apoio, recrutamento, promoção ou ajuda a organizações criminosas ou terroristas ou a seus atos, entre outros[286].

A MP também foi criticada e considerada inconstitucional por especialistas, pela ausência de urgência em sua edição, especialmente por já estar em discussão no Congresso Nacional o Projeto de Lei ("PL") nº 2.630/20 ("PL das fake news"). Ademais, a MP impedia o combate à desinformação, ao *bullying*, e, "dependendo de como os conceitos de 'segurança pública', 'defesa nacional' e 'segurança do Estado' fossem operacionalizados, a ataques antidemocráticos"[287]. Isso porque a remoção desse tipo de conteúdo, pela MP, dependeria de decisão judicial prévia, não só indo na contramão das propostas reguladoras mundiais, mas violando ainda

[286] "Art. 8º-C Em observância à liberdade de expressão, comunicação e manifestação de pensamento, a exclusão, a suspensão ou o bloqueio da divulgação de conteúdo gerado por usuário somente poderá ser realizado com justa causa e motivação.
§ 1º Considera-se caracterizada a justa causa nas seguintes hipóteses:
I - quando o conteúdo publicado pelo usuário estiver em desacordo com o disposto na Lei nº 8.069, de 13 de julho de 1990;
II - quando a divulgação ou a reprodução configurar:
a) nudez ou representações explícitas ou implícitas de atos sexuais;
b) prática, apoio, promoção ou incitação de crimes contra a vida, pedofilia, terrorismo, tráfico ou quaisquer outras infrações penais sujeitas à ação penal pública incondicionada;
c) apoio, recrutamento, promoção ou ajuda a organizações criminosas ou terroristas ou a seus atos;
d) prática, apoio, promoção ou incitação de atos de ameaça ou violência, inclusive por razões de discriminação ou preconceito de raça, cor, sexo, etnia, religião ou orientação sexual;
e) promoção, ensino, incentivo ou apologia à fabricação ou ao consumo, explícito ou implícito, de drogas ilícitas;
f) prática, apoio, promoção ou incitação de atos de violência contra animais;
g) utilização ou ensino do uso de computadores ou tecnologia da informação com o objetivo de roubar credenciais, invadir sistemas, comprometer dados pessoais ou causar danos a terceiros;
h) prática, apoio, promoção ou incitação de atos contra a segurança pública, defesa nacional ou segurança do Estado;
i) utilização ou ensino do uso de aplicações de internet, sítios eletrônicos ou tecnologia da informação com o objetivo de violar patente, marca registrada, direito autoral ou outros direitos de propriedade intelectual;
j) infração às normas editadas pelo Conselho Nacional de Auto-regulamentação Publicitária referentes a conteúdo ou material publicitário ou propagandístico;
k) disseminação de vírus de software ou qualquer outro código de computador, arquivo ou programa projetado para interromper, destruir ou limitar a funcionalidade de qualquer recurso de computador; ou
l) comercialização de produtos impróprios ao consumo, nos termos do disposto no § 6º do art. 18 da Lei nº 8.078, de 11 de setembro de 1990;
III - requerimento do ofendido, de seu representante legal ou de seus herdeiros, na hipótese de violação à intimidade, à privacidade, à imagem, à honra, à proteção de seus dados pessoais ou à propriedade intelectual; ou;
IV - cumprimento de determinação judicial". Disponível em <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/Mpv/mpv1068.htm>. Acesso em: 34 nov. 2022.

[287] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 152.

PUC-Rio - Certificação Digital Nº 2112697/CA

a livre iniciativa das plataformas, que têm o direito de definir o tipo de ambiente que querem fornecer aos seus usuários em seus próprios termos de uso.

Entende-se, ainda, que a MP violava a liberdade de expressão dos usuários, que ficariam presos em ambientes digitais tóxicos, repletos de discurso ofensivo e fraudulento, já que não é possível, na prática, que o Poder Judiciário consiga controlar esse tipo de discurso em tempo hábil para evitar sua proliferação[288].

Por fim, considera-se que a MP agiu mal ao atribuir a uma autoridade administrativa a competência para aplicar sanções por eventuais violações aos seus termos, pois instituiu, assim, espécie de censura administrativa sobre o conteúdo publicado nas redes sociais[289].

Por essas razões, agiu bem o Senado Federal ao devolver a MP, destacando que ela versava sobre o mesmo tema do PL das *fake news*, já em discussão no Congresso Nacional. Ademais, a Ministra Rosa Weber, relatora de sete ações diretas de inconstitucionalidade propostas contra a MP, também suspendeu na íntegra sua eficácia, destacando que:

> [A] propagação de fake news, de discurso de ódio, de ataques às instituições e à própria democracia, bem como a regulamentação da retirada de conteúdos de redes sociais, consubstanciam um dos maiores desafios contemporâneos à conformação dos direitos fundamentais. Não por outra razão, este Supremo Tribunal Federal, o Tribunal Superior Eleitoral e o Congresso Nacional têm enfrentado, cada um dentro de suas competências constitucionais[290].

O PL das *fake news* (nº 2.630/20), iniciado em maio de 2020, institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na internet. A versão final aprovada no Senado em junho de 2020 estabelece normas sobre transparência para provedores de redes sociais e serviços de mensageria privada com dois milhões ou mais de usuários registrados no Brasil. O PL tem compromisso com a liberdade de expressão e a pluralidade, pois contempla princípios como a liberdade de expressão e de impressa, a responsabilidade compartilhada pela preservação de uma esfera pública livre, plural, diversa e democrática. Consagra ainda como princípio o fomento à livre formação de preferências políticas e de visão de mundo pessoal de usuários e a necessidade de se garantir transparência nas regras para veiculação

[288] MENDONÇA, Eduardo. *Retrocesso autoritário*. Estadão, jun. 2021. Disponível em: <https://politica.estadao.com.br/blogs/fausto-macedo/retrocesso-autoritario/>. Acesso em: 10 out. 2022.
[289] Ibid.
[290] BRASIL. STF. ADI nº 6.991, Rel. Min. Rosa Weber, decisão monocrática, j. em 14/09/2021,. Disponível em: <https://www.jusbrasil.com.br/jurisprudencia/stf/1284309785/inteiro-teor-1284309815>. Acesso em: 10 out. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

de anúncios e conteúdo pago. Por fim, o PL contempla direitos da personalidade, dignidade, honra e privacidade, que podem ser fundamentos para limitar a liberdade de expressão, desde que de forma proporcional.

O PL não entra no mérito de tentar definir discursos ilícitos e atribuir ao Estado o controle sobre o que é publicado na internet. Os principais objetivos do PL são: (i) o fortalecimento do processo democrático por meio do combate à desinformação e do fomento à diversidade de informações na internet no Brasil; (ii) a busca por maior transparência sobre conteúdos pagos disponibilizados para o usuário; e (iii) desencorajar o uso de contas inautênticas para disseminar desinformação nas aplicações de internet[291]. Ocorre que o PL tem deficiências importantes que impedem a ampla concretização de seus objetivos e princípios.

Em primeiro lugar, entende-se que o PL agiu bem em dar foco ao comportamento, e não ao conteúdo, ao, por exemplo, prever que as plataformas adotem medidas para vedar: (i) contas inautênticas; (ii) disseminadores artificiais não rotulados, (iii) redes de disseminação artificial que disseminem desinformação; e (iv) conteúdos patrocinados não rotulados. Ao assim agir, o PL permite o uso de critérios objetivos (como a existência de contas falsas ou inautênticas), que podem ser aplicados de forma imparcial pelas plataformas. Para isso, contudo, os termos de uso das plataformas devem esclarecer os critérios adotados para classificar uma conta como inautêntica ou falsa e o que caracteriza comportamento coordenado inautêntico e como ele é identificado. Por isso, entende-se positiva a previsão do PL de que os relatórios de transparência indiquem o número total de contas automatizadas e de redes de distribuição artificial detectadas, com a especificação das medidas adotadas e suas motivações, bem como a metodologia utilizada na detecção da irregularidade[292].

No entanto, para dar efetividade a essa previsão, é preciso exigir que o detalhamento sobre a metodologia utilizada na detecção da irregularidade inclua informações específicas, como: (i) quantas contas constituem uma rede que enseja a desativação de todas as contas envolvidas, mesmo quando algumas sejam autênticas?; (ii) quais os critérios adotados para classificar uma conta como falsa

[291] Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&ts=1648639813988&disposition=inline>. Acesso em: 10 out. 2022.

[292] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 160.

PUC-Rio - Certificação Digital Nº 2112697/CA

ou inautêntica?; (iii) alegações falsas sobre o nome ou localização são suficientes para que uma conta seja considerada inautêntica?; (iv) o que significa comportamento coordenado e como ele é identificado pelas plataformas?

Em segundo lugar, em relação aos procedimentos de moderação de conteúdo, o PL prevê que as plataformas devem disponibilizar "mecanismos de recurso e devido processo", em seu artigo 12, incluindo o direito do usuário de recorrer da indisponibilização de conteúdos e contas, em seu artigo 12, §3º. Também prevê que, em caso de denúncia ou de medida aplicada em função dos termos de uso ou do PL, o usuário deve ser notificado sobre a fundamentação, o processo de análise, e a aplicação da medida, assim como sobre os prazos e procedimentos para sua contestação. Nesse ponto, contudo, considerando que muitas decisões partem do uso de ferramentas automatizadas, elas devem ser consideradas fundamentadas para fins de cumprir o dispositivo do PL sempre que apontarem a previsão dos termos de uso violada, sem que se exija maiores detalhamentos do caso concreto, o que pode inviabilizar a moderação de conteúdo em escala e, assim, permitir a ampla disseminação de conteúdo tóxico[293].

O § 4º do artigo 12 prevê que, havendo dano decorrente da caracterização equivocada de conteúdo como violador dos termos de uso ou da Lei, caberá ao provedor da rede social repará-lo, no âmbito e nos limites técnicos de seu serviço. Esse ponto é problemático, pois, além de ser contrário à tendência mundial, desconsidera dois pontos fundamentais da moderação de conteúdo. O primeiro é que, no campo da liberdade de expressão, a avaliação sobre aplicação devida ou não dos termos de uso ou de disposições legais sobre um conteúdo será sempre subjetiva. Por isso, a previsão pode gerar responsabilização indevida ou criar incentivos inadequados para que as plataformas evitem moderar conteúdo de determinados temas, como desinformação, para evitar responsabilização[294].

O segundo é que, como a moderação de conteúdo ocorre sempre em escala, com o uso de ferramentas automatizadas, haverá sempre decisões equivocadas. Por isso, a responsabilização das plataformas não deve ter por base casos específicos de erros, que são inevitáveis, mas uma avaliação global do funcionamento do sistema

[293] Ibid, p. 162.
[294] Ibid, p. 163.

PUC-Rio - Certificação Digital Nº 2112697/CA

de moderação de conteúdo[295]. Além disso, a apuração da extensão dos danos causados com a restrição de um conteúdo não é algo objetivamente verificável.

Em terceiro lugar, o PL agiu bem ao estabelecer requisitos de transparência, exigindo a publicação de relatórios trimestrais para informar sobre os procedimentos e as decisões de tratamento de conteúdo gerado por terceiros, e sobre as medidas adotadas para cumprimento da Lei. Os relatórios devem conter, no mínimo:

> I - número de com contas registrada em solo brasileiro na plataforma e número de usuários brasileiros ativos no período analisado; II - número de contas inautênticas removidas da rede, com classificação do comportamento inautêntico, incluindo a porcentagem de quantas estavam ativas; III - número de disseminadores artificiais, conteúdos, conteúdos patrocinados não registrados no provedor de aplicações que foram removidos da rede ou tiveram o alcance reduzido, com classificação do tipo de comportamento inautêntico e número de visualizações; IV - número de reclamações recebidas sobre comportamento ilegal e inautêntico e verificações emitidas no período do relatório, indicando a origem e o motivo da reclamação; V - tempo entre o recebimento das reclamações pelo provedor de aplicação e a resposta dada, discriminado de acordo com o prazo para resolução da demanda; VI – dados relacionados a engajamentos ou interações com conteúdos que foram verificados como desinformação (...); VII - estrutura dedicada ao combate à desinformação no Brasil, em comparação a outros países, contendo o número de pessoal diretamente empregado na análise de conteúdo bem como outros aspectos relevantes; VIII - em relação a conteúdo patrocinado, quem pagou pelo conteúdo, qual o público alvo e quanto foi gasto, em uma plataforma de fácil acesso a usuários e pesquisadores[296].

Embora o PL caminhe na direção certa ao revelar uma preocupação com a garantia de maior transparência, percebe-se que as informações exigidas não são suficientes para promover uma *accountability* adequada. Como sustenta Luna Barroso:

> As plataformas devem ter obrigações adicionais, inclusive de disponibilização do conteúdo que foi objeto de algum tipo de análise pela plataforma em uma base de dados acessível, ao menos, a pesquisadores, para que se possa avaliar o mérito das decisões de moderação de conteúdo. Sem acesso ao conteúdo de fundo, apenas aos números, não é possível avaliar se os termos e condições são aplicados como dito ou mesmo se estão sendo aplicados de forma isonômica, e pode criar incentivos para a remoção em excesso, na tentativa de mostrar números elevados de remoção[297].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[295] Ibid.

[296] Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra;jsessionid=node0gvtn9f3zjslha689218o3lpc12494207.node0?codteor=1909983&filename=PL+2630/2020>. Acesso em: 10 out. 2022.

[297] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 164.

Em quarto lugar, o PL prevê a criação do Conselho de Transparência e Responsabilidade na Internet, a ser instituído pelo Congresso Nacional. O Conselho será responsável pelo acompanhamento das medidas que trata e outras funções, previstas no § único do artigo 25. O artigo 26 dispõe sobre a composição do Conselho, que incluirá representantes da Câmara, Senado, Conselho Nacional de Justiça, do Comitê Gestor da Internet no Brasil ("CGI.br"), representantes da sociedade civil, da academia e comunidade técnica, de provedores de acesso, aplicações e conteúdo, dentre outros. As despesas do Conselho correrão à conta do orçamento do Senado Federal.

Nesse ponto, embora se concorde com a criação de um órgão especializado para acompanhamento das exigências do PL, não ficou claro no texto se sua competência seria limitada a uma função de consulta, para que o Poder Judiciário dê a palavra final sobre eventual aplicação de sanções, ou se a fiscalização e aplicação das sanções estaria concentrada exclusivamente no Conselho. Além disso, a previsão de que o código de conduta dos intermediários seja avaliado e aprovado pelo Congresso Nacional engessa o funcionamento e flexibilidade do órgão, além de reduzir sua independência. Ainda, a competência do órgão para avaliar a adequação das políticas de uso adotadas pelas plataformas pode representar intervenção excessiva sobre uma atividade privada, que deveria ser limitada à confirmação de que os termos e condições têm previsões para endereçar conteúdo ilícito ou danoso. Por fim, a proposta pode apresentar custos para criação de novo órgão, que terá estrutura semelhante ao CGI.br, que poderia exercer a função do Conselho, sem a necessidade de criação de novo órgão vinculado ao Congresso e financiado pelo Senado[298].

Em quinto lugar, o PL prevê, em seu artigo 30, que as plataformas poderão criar instituições de autorregulação voltadas à transparência e à responsabilidade no uso da internet, que deverá ser certificada pelo Conselho de Transparência e Responsabilidade. A instituição de autorregulação poderá elaborar relatórios trimestrais e informações acerca das políticas de uso e de monitoramento de volume de conteúdo compartilhado pelos usuários[299]. A previsão, contudo, é confusa. Seu

[298] Ibid, p. 166.

[299] "Art. 30. Os provedores de redes sociais e de serviços de mensageria privada poderão criar instituição de autorregulação voltada à transparência e à responsabilidade no uso da internet, com as seguintes atribuições: I – criar e administrar plataforma digital voltada à transparência e à responsabilidade no uso da internet, que contenha regras e procedimentos para decidir sobre a

PUC-Rio - Certificação Digital Nº 2112697/CA

título sugere a criação de um modelo de "autorregulação regulada", mas a proposta detalhada é apenas de uma instituição de autorregulação, que elabora relatórios autônomos que podem ou não ser enviados ou considerados pelo Conselho. O PL, na forma atual, não delimita de forma clara qual seria o papel do Estado, qual seria o papel do Conselho, e qual seria o papel do órgão de "regulação regulada" na implementação da lei, podendo gerar confusão sobre quem tem competência efetiva para fiscalizar a aplicar sanção por seu descumprimento[300].

Por fim, em relação às sanções, que serão aplicadas pelo Poder Judiciário, o PL prevê que:

> Art. 31. Sem prejuízo das demais sanções civis, criminais ou administrativas, os provedores de redes sociais e de serviços de mensageria privada ficam sujeitos a:
> I – advertência, com indicação de prazo para adoção de medidas corretivas; ou
> II – multa de até 10% (dez por cento) do faturamento do grupo econômico no Brasil no seu último exercício.
> § 1º Na aplicação da sanção, a autoridade judicial observará a proporcionalidade, considerando a condição econômica do infrator, as consequências da infração na esfera coletiva e a reincidência.
> § 2º Para os efeitos desta Lei, será considerado reincidente aquele que repetir, no prazo de 6 (seis) meses, condutas anteriormente sancionadas.

No entanto, além do fato de que a responsabilização das plataformas por violação ao sistema de autorregulação regulada deveria se dar apenas a partir de uma análise sistêmica, e não por violações pontuais, fato é que o Judiciário não é um órgão especializado para tomar decisões que levem em consideração a situação real do setor. Como pontuam Juliano Maranhão e Ricardo Campos:

> O Poder Judiciário não possui *expertise* e velocidade necessárias para a reação eficiente contra a produção e divulgação de fake news, e, por atuar caso a caso, não

PUC-Rio - Certificação Digital Nº 2112697/CA

adoção de medida informativa, atendendo ao disposto nesta Lei; II – assegurar a independência e a especialidade de seus analistas; III – disponibilizar serviço eficiente de atendimento e encaminhamento de reclamações; IV – estabelecer requisitos claros, objetivos e acessíveis para a participação dos provedores de redes sociais e serviços de mensageria privada; V – incluir em seu quadro uma ouvidoria independente com a finalidade de receber críticas e avaliar as atividades da instituição; e VI – desenvolver, em articulação com as empresas de telefonia móvel, boas práticas para suspensão das contas de usuários cuja autenticidade for questionada ou cuja inautenticidade for estabelecida. § 1º A instituição de autorregulação deverá ser certificada pelo Conselho de Transparência e Responsabilidade na Internet. § 2º A instituição de autorregulação poderá elaborar e encaminhar ao Conselho de Transparência e Responsabilidade na Internet relatórios trimestrais em atendimento ao disposto nesta Lei, bem como informações acerca das políticas de uso e de monitoramento de volume de conteúdo compartilhado pelos usuários dos serviços de mensageria privada. § 3º A instituição de autorregulação aprovará resoluções e súmulas de modo a regular seus procedimentos de análise." Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra;jsessionid=node0gvtn9f3zjslha689218o3lpc12494207.node0?codteor=1909983&filename=PL+2630/2020>. Acesso em: 10 nov. 2022.

[300] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 167.

> tem ferramentas institucionais para identificar e combater organizações que estejam por trás de disseminação massiva e articulada de notícias fraudulentas[301].

Por isso, a função deveria pertencer a um órgão especializado e independente[302].

Na Câmara dos Deputados foi apresentado, em outubro de 2021, um substitutivo ao PL pelo seu relator na Câmara, o deputado Orlando Silva, e aguarda-se nesse momento a criação de Comissão Temporária pela MESA para aprovação do mesmo.

Em primeiro lugar, diferente da versão original, o substitutivo da Câmara pretende regular também ferramentas de busca, e que a lei seja aplicável apenas para plataformas com mais de 10 milhões de usuários registrados no Brasil, e não 2 milhões, como propunha a versão do Senado, passando a ter como foco apenas as plataformas mais populares.

O substitutivo também altera a seção IV do PL original, intitulada "Dos procedimentos de moderação", para uma seção intitulada "Dos procedimentos do devido processo legal". A seção foi quase toda alterada, prevendo que, ao aplicar regras próprias que impliquem na exclusão, indisponibilização, redução de alcance ou sinalização de conteúdo, as plataformas devem notificar os usuários sobre: a) a natureza da medida aplicada; b) a fundamentação, que deve necessariamente apontar a cláusula aplicada de suas regras e o conteúdo ou a conta que deu causa à decisão; c) os procedimentos e prazos para exercer o direito de pedir a revisão da decisão; e d) se a decisão foi tomada exclusivamente por meio de sistemas automatizados, fornecendo informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão.

No substitutivo, a Câmara dos Deputados também propõe que as plataformas disponibilizem canal próprio, destacado e de fácil acesso, para consulta das informações prestadas, formulação de denúncias sobre conteúdos e contas, bem como envio de pedido de revisão de decisões. Quanto à reparação de eventuais danos causados, o substitutivo prevê que a autoridade judicial pode determinar aos provedores uma reparação, que consiste no envio de informações a todos os

[301] MARANHÃO, Juliano; CAMPOS, Ricardo. Fake News e autorregulação regulada das redes sociais no Brasil: fundamentos constitucionais. *In:* ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. *Fake News e Regulação*. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021, p. 343.

[302] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 168.

PUC-Rio - Certificação Digital Nº 2112697/CA

impactados pelo conteúdo problemático, com o mesmo alcance do conteúdo inadequado.

O substitutivo também acrescenta um crime em espécie, qual seja, o de promover, constituir, financiar ou integrar ação coordenada, mediante uso de contas automatizadas (robôs) e outros meios ou expedientes não fornecidos diretamente pelo provedor de aplicação de internet. De acordo com o relatório da Câmara dos Deputados, esse crime se aplica ao disparo em massa de mensagens que veiculem conteúdo passível de sanção criminal ou fatos comprovadamente inverídicos capazes de colocar em risco a vida, integridade física e mental, segurança das pessoas, além da higidez do processo eleitoral. A pena proposta é de reclusão de 1 a 5 anos e multa.

De forma positiva, o substitutivo exclui a previsão de criação do Conselho de Transparência e Responsabilidade na Internet, atribuindo suas funções ao CGI.br, órgão já existente, economizando tempo e recursos. Também afasta o controle do Congresso Nacional sobre códigos de conduta a serem adotados pelas plataformas, e afasta o controle do CGI.br sobre as políticas de uso adotadas pelas plataformas.

Em relação ao sistema de autorregulação regulada, o substitutivo estabelece que provedores deverão (e não mais poderão) criá-los, com as seguintes funções:

> I - criar e administrar plataforma digital voltada ao recebimento de denúncias sobre conteúdo ou contas e tomada de decisão sobre medidas a serem implementadas por seus associados, bem como a revisão de decisões de conteúdos e contas, por meio de provocação por aqueles afetados diretamente pela decisão; II - tomar decisões, em tempo útil e eficaz, sobre as denúncias e revisão de medidas abrangidos por esta lei; III – desenvolver, em articulação com as empresas de telefonia móvel, boas práticas para suspensão das contas de usuários cuja autenticidade for questionada ou cuja inautenticidade for estabelecida; e IV – estabelecer e divulgar em seu sítio na internet Código de Conduta para a implementação desta Lei, vinculante para seus associados, resoluções sobre seus procedimentos de análises e súmulas interpretativas, com base na experiência de seu órgão decisório.

Revela-se, contudo, preocupante que decisões sobre moderação de conteúdo sejam tomadas de forma unificada por todas as plataformas. Para garantir o objetivo de pluralidade do PL, é normal e até desejável que plataformas tenham algum grau de liberdade para aplicarem seus termos de uso e chegarem a conclusões diversas sobre a licitude ou ilicitude de determinado discurso[303].

[303] Ibid, p. 173.

PUC-Rio - Certificação Digital Nº 2112697/CA

Embora o substitutivo apresente alterações positivas, a proposta continua a atribuir aos tribunais o papel de fiscalização e aplicação de sanções. Entende-se, contudo, que um modelo mais eficiente de autorregulação regulada ou corregulação atribuiria a um único órgão independente o poder de fiscalizar o cumprimento da lei, como se verá no Capítulo 3.3[304].

PUC-Rio - Certificação Digital Nº 2112697/CA

[304] Ibid, p. 173.

Capítulo 3 – A perspectiva procedimental da regulação estatal da moderação de conteúdo e a garantia da liberdade de expressão dos usuários

Devido à influência das grandes plataformas de redes sociais no discurso público, os governos e a sociedade civil passaram a exigir ações regulatórias quanto à atividade de moderação de conteúdo.

A preocupação com os novos poderes exercidos por empresas privadas consolidou o conceito de constitucionalismo digital[305]. Por esse conceito, defende-se um marco teórico que renove os dois objetivos tradicionais do constitucionalismo – a garantia de direitos e a limitação de poderes – com base em uma nova realidade em que o poder é compartilhado por Estados e por grandes empresas privadas[306].

Para Nicolas Suzor, o constitucionalismo digital enfrenta o desafio de regular poderes que estão distribuídos entre muito atores, dentro de sistemas complexos, com diversos componentes de interação. Para os governos, esse novo modelo demanda que se repense a forma de regulação que pode operar em um ambiente descentralizado, ou seja, onde o Estado não é o único, ou mesmo o mais poderoso ator que procura regular comportamentos[307].

Por isso mesmo, é preciso pensar, para além de respostas normativas tradicionais, em instrumentos jurídicos inovadores que surgem nesse novo contexto transacional[308]. Esses novos instrumentos podem ser pensados além das dimensões tradicionais que são construídas em torno de Estados (como ordenamentos jurídicos nacionais ou organizações regionais/internacionais), como no âmbito de regras autônomas dos atores privados globais, representados nesse estudo pelas plataformas de redes sociais[309].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[305] "Constitucionalismo digital é um conceito que se refere a um contexto específico, o ambiente digital, no qual atores privados emergem ao lado de estados-nação como potenciais violadores de direitos fundamentais. Essa particularidade do ambiente digital requer que o conceito de constitucionalismo se liberte da dimensão estatal de modo a dimensionar adequadamente a emergência dos poderes de atores privados". CELESTE, Edoardo. *Digital Constitutionalism*: mapping the constitutional responses to digital technology's challenges. HIIG Discussion Paper Series No. 2018-02 (2018).

[306] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 130.

[307] SUZOR, Nicolas. *Lawless:* the secret rules that govern or digital lives, Cambridge University Press, 2019, p. 107-108.

[308] CELESTE, Edoardo. Op. cit., p. 17-19.

[309] NITRINI, Rodrigo Vidal. Op. cit., p. 131.

Nesse contexto, Nicolas Suzor fornece dois principais caminhos para o processo de constitucionalização das plataformas de redes sociais. Em primeiro lugar, entende que as grandes plataformas atuam hoje, em grande parte, à margem do direito, pois têm ampla discricionariedade para criar e aplicar suas próprias regras, da forma como preferirem, como pouca *accountability* sobre suas decisões. Isso permite que as decisões sejam arbitrárias, imprevisíveis e inconsistentes, sendo o oposto da maneira jurídica de tomada de decisões[310].

Por isso, o autor entende que é preciso aplicar garantias análogas às do "*rule of law*"[311] às grandes plataformas. Esse enfoque se traduz na necessidade de imposição de regras claras e bem-definidas pelas plataformas, com decisões motivadas. Apenas a partir daí será possível pensar em um devido processo digital. Assim, garantias procedimentais se mostram necessárias para conciliar as preocupações do constitucionalismo digital com algum grau de autonomia que as plataformas têm (e devem ter) para criação de suas regras[312].

A garantia de um processo de constitucionalização não se confunde, contudo, com posições substantivas rígidas. Para Nicolas Suzor, garantias processuais e procedimentais são necessárias, mas devem vir acompanhadas de direitos materiais. Por isso, em segundo lugar, o autor defende a incorporação, pelas plataformas, de uma lógica de proteção aos direitos humanos. O autor não advoga pela adição de critérios substantivos, mas sim por uma lógica de respeito aos direitos fundamentais, a partir dos direitos humanos[313]. Isso significa que restrições à liberdade de expressão devem ser devidamente fundamentadas, e que a articulação dessas razões pode se valer do repertório jurídico dos direitos humanos[314].

David Kaye também entende que as plataformas de redes sociais têm muito poder para decidir o que pode ou não ser publicado, o que coloca em risco a liberdade de expressão. Por isso, para o autor também é necessário usar parâmetros

PUC-Rio - Certificação Digital Nº 2112697/CA

[310] SUZOR, Nicolas. *Lawless:* the secret rules that govern or digital lives, Cambridge University Press, 2019, p. 106-110.
[311] Para o autor, a principal ideia do "rule of law" é que as pessoas saibam as razões pelas quais as decisões que as afetam são tomadas, por meio de regras previamente definidas e aplicadas de modo isonômico. Ibid.
[312] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 133.
[313] SUZOR, Nicolas. *Lawless:* Op. cit., p. 130-131.
[314] NITRINI, Rodrigo Vidal. Op. cit., p. 135.

de direitos humanos na avaliação do conteúdo, para evitar a exploração tanto do controle estatal quanto do direito privado[315].

Traçando um paralelo com autores brasileiros, Aline Osório, em trabalho voltado ao direito eleitoral e liberdade de expressão, defende que mandamentos contidos em tratados internacionais sobre direitos humanos dos quais o Brasil é signatário, ainda que não internacionalizados, sejam levados em consideração na interpretação constitucional, "para que se adote uma compreensão adequada da liberdade de expressão e do sistema constitucional brasileiro de tutela desse direito"[316].

Em que pesem as propostas dos autores, é necessário destacar que pensar na adoção de parâmetros dos direitos humanos não dá fim aos dilemas inerentes aos grandes sistemas de moderação de conteúdo. Afinal, há divergências interpretativas sobre a aplicação dos direitos envolvidos entre diversos órgãos internacionais, e a própria natureza interpretativa de direitos fundamentais pode levar a divergências em torno de suas aplicações concretas. Além disso, são parâmetros voltados a Estados e que não devem ser transferidos para plataformas privadas[317].

Rodrigo Vidal Nitrini entende que as propostas de Nicolas Suzor e David Kaye em favor de uma incorporação de direitos humanos pelas plataformas são pertinentes, mas devem ser exploradas e qualificadas, pois, embora a ênfase dos autores seja decisões de moderação de conteúdo, suas considerações acabam abarcando sugestões mais abrangentes, como a adoção, pelas plataformas, de mecanismos internos para avaliação de impactos de suas políticas sobre direitos humanos nos mercados em que atuam. De todo modo, em relação especificamente à moderação de conteúdo, a proposta de ambos é a incorporação de uma lógica de direitos, pela qual restrições a direitos fundamentais por plataformas de redes sociais devem ser motivadas. Com efeito, a motivação sobre as regras e decisões tomadas pelas plataformas permite um maior controle social, e viabiliza uma responsabilização jurídica pelo Poder Judiciário[318].

PUC-Rio - Certificação Digital Nº 2112697/CA

[315] KAYE, David. *Speech Police:* the global struggle to govern the internet, Columbia Global Reports, 2019, p. 118-121.
[316] OSORIO, Aline. *Direito eleitoral e liberdade de expressão*. Belo Horizonte: Fórum, 2017, p. 48.
[317] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 138.
[318] Ibid, p. 137.

Como os parâmetros substantivos de direitos fundamentais estão sempre em aberto, pois não são definidos previamente, é necessário consolidar um devido processo digital junto às plataformas.

3.1 Liberdade editorial das plataformas de redes sociais como novos agentes na estruturação do discurso público

Como visto ao longo deste trabalho e das conclusões já apresentadas, as plataformas não são neutras. Com efeito, cada plataforma de rede social tem seus termos de uso e diretrizes de comunidade feitos com base em sua própria percepção sobre o que querem que esteja presente em seu ambiente virtual, e tomam as decisões sobre moderação de conteúdo com base nessas políticas internas. Por isso, como as plataformas não são neutras, é possível afirmar que há hoje uma nova espécie de liberdade editorial.

Essa nova espécie de liberdade editorial é fundamental para compreender o novo papel que as plataformas de redes sociais exercem na regulação do discurso digital, no que pode ou não ser dito, no que deve ou não permanecer *online*. Kate Klonick e Thomas Kadri sumarizam esse novo papel exercido pelas plataformas de redes sociais:

> Na época dominada pelos velhos governantes, a governança de discursos era essencialmente dividida entre o legislativo, o executivo, o judiciário e a imprensa (...). As decisões da imprensa a respeito de o que publicar - o que possuía interesse noticioso - eram feitas por conselhos editoriais a quem era dada alguma deferência pelas cortes (...). Hoje, na era dos novos governantes, nós podemos ver sombras desses vários papéis, mas em uma construção razoavelmente diversa. Grande parte da governança do discurso online é feita por plataformas privadas que exercem todos esses papéis - legislativo, executivo, judiciário e de imprensa - todos de uma vez[319].

Com efeito, só é possível pensar na criação de regras de moderação de conteúdo quando essa atividade é feita em conjunto com outros aspectos da governança privada, como a fixação das condições de publicação ou da curadoria algorítmica que define quais conteúdos serão exibidos para quem, e em qual prioridade. Esse novo tipo de julgamento editorial é facilmente perceptível quando se observa as decisões de uma plataforma que determinam a manutenção de uma

[319] KADRI, Thomas e KLONICK, Kate. *Facebook v. Sullivan*: public figures and newsworthiness in online speech, Southern California Lay Review Volume 93 (2019), p. 94.

PUC-Rio - Certificação Digital Nº 2112697/CA

postagem no ar em razão de ser de interesse público, ou que determinam a remoção de determinado conteúdo por conter informação falsa, por exemplo[320].

Esse novo tipo de liberdade editorial veio à tona para definir a atividade de moderação de conteúdo exercida pelas plataformas de redes sociais, pois essa atividade não se enquadra nas funções dos "velhos governantes"[321]. Com efeito, diferentemente dos "velhos governantes", as plataformas de redes sociais não exercem uma curadoria específica sobre o que pode ou não ser publicado, pois essa lógica, geralmente, é feita de forma inversa pelas redes sociais, ou seja, após uma publicação ser feita pelos usuários. Existe intervenção editorial, mas aqui essa intervenção é feita de forma excepcional, não influenciando no conteúdo postado na plataforma, que, como regra, trabalha com um espaço de publicação livre[322].

Como as plataformas não decidem o que é postado, ou seja, não exercem o papel dos editores tradicionais (os "velhos governantes"), não se pode pensar nas plataformas com os mesmos deveres e obrigações que exercem canais de mídias tradicionais, como jornais e televisão. As diferentes funções editoriais exercidas pelas plataformas de redes sociais demanda tratamento jurídico distinto a elas, como corretamente fez o artigo 19 do Marco Civil da Internet, ao isentá-las de responsabilidade civil pelo conteúdo postado por terceiros, salvo em caso de descumprimento de ordem judicial específica.

No entanto, o mero fato de existir liberdade editorial às plataformas, ainda que exercida *a posteriori* e em novas formas, implica necessariamente em escolhas valorativas das empresas, fruto de sua autonomia[323].

Cass R. Sunstein entende que, embora as plataformas de redes sociais não decidam o que é ou não postado, elas decidem como o conteúdo é entregue aos usuários. O autor esclarece que a arquitetura das redes sociais não privilegia as descobertas feitas por acaso (*serendipity*). Diferentemente do que acontece quando se abre um jornal ou se cruza os espaços públicos da cidade, nas redes sociais os cidadãos estão confinados em bolhas e expostos a conteúdos que apenas reforçam suas próprias convicções e interesses. E isso porque as redes sociais, com base em

PUC-Rio - Certificação Digital Nº 2112697/CA

[320] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 120.
[321] KLONICK, Kate. "*The New Governors: The People, Rules, and Processes Governing Online Speech*," Harvard Law Review 131 (2018): 1598.
[322] NITRINI, Rodrigo Vidal. op. cit., p. 121.
[323] Ibid, p. 123.

dados pessoais do usuário, focam na personalização e segmentação de conteúdos, com a ideia de soberania do interesse do consumidor, deixando, assim, de servir aos interesses da coletividade.

> A preocupação não é que as coisas estejam piorando. É que o aumento da capacidade tecnológica de autoclassificação e de personalização está criando sérios problemas. O que as plataformas de mídia social fazem é tornar certos tipos de segmentação e certos tipos de autoclassificação, e especialmente autoclassificação entre centenas, milhares ou milhões de estranhos muito mais fáceis – mais fáceis do que nunca. Já tivemos muitas câmaras de eco e segmentação antes, mas segmentar pessoas com maior probabilidade de acreditar em falsidades específicas e câmaras de eco de um clique são algo novo.
> (...)
> a utilização do Facebook, Twitter, Instagram e outros depende de uma variedade de fatores, incluindo escolhas individuais, algoritmos relevantes, normas sociais e julgamentos arquitetônicos dos próprios designers da plataforma. Um feed de notícias, ou algo parecido, pode promover casulos de informações ou desencorajá-los. As plataformas podem fornecer salvaguardas no caso de processos democráticos serem intencionalmente interrompidos ou falsidades prejudiciais se espalharem; pode ajudar as pessoas a descobrir o que é verdadeiro. (Ultimamente, o Facebook tem feito exatamente isso)[324].

Vê-se, portanto que existe uma zona cinzenta entre o controle editorial realizado pelos "velhos governantes" e aquele feito pelas plataformas de redes sociais, antes mesmo de realizarem a moderação de conteúdo, ou seja, ao fazerem o conselho editorial do *feed* de cada usuário.

Além disso, essa nova forma de liberdade editorial das plataformas permite esclarecer o porquê de algumas decisões autônomas dessas empresas privadas permitirem solucionar problemas reais e concretos. O fato de o Facebook ter ocultado informações falsas sobre vacinas durante a pandemia do COVID-19 não significa que a empresa deve se engajar sempre para separar postagens verdadeiras

PUC-Rio - Certificação Digital Nº 2112697/CA

[324] Tradução livre de: "The concern is not that things are getting worse. It is that the increased technological capacity for self-sorting and for personalization is creating serious problems. What social media platforms do is to make certain kinds of targeting and certain kinds of self-sorting, and especially self-sorting among hundreds, thousands, or millions of strangers a lot easier — easier than it has ever been. We have had plenty of targeting and echo chambers before, but targeting people who are especially likely to believe specific falsehoods, and one-click echo chambers, are something new. (…)
The uses of Facebook, Twitter, Instagram, and others depends on a variety of factors, including individual choices, relevant algorithms, social norms, and the architectural judgments of the platform designers themselves. A News Feed, or anything like it, can promote information cocoons, or discourage them. Platforms can provide safeguards in the event that democratic processes are being intentionally disrupted or harmful falsehoods are spreading; it can help people find out what is true. (Of late, Facebook has been doing exactly that.). Disponível em: <https://about.fb.com/news/2018/01/sunstein-democracy/>. Acesso em: 02 jan. 2023.

de falsas. Eventuais restrições podem ser justificadas por conta de riscos concretos que se apresentam em cada plataforma[325].

Rodrigo Nitrini traça um paralelo entre essa nova liberdade editorial e a derrubada de postagens do então presidente da república brasileiro e do presidente venezuelano Nicolás Maduro, em que se defendia o uso do medicamento cloroquina e, respectivamente, de uma receita caseira para combater o vírus, durante a pandemia, pelo Facebook e outras plataformas:

> Essas derrubadas, no entanto, não significam um silenciamento completo dos discursos políticos desses dois presidentes, que continuam possuindo amplos meios, inclusive oficiais, para divulgarem suas ideias. Ainda assim, caracterizam essa disposição das grandes redes sociais de realizarem um nível de controle editorial dentro de seus ambientes, que não se confundem com a internet em geral, diante de riscos que são próprios de suas atividades (tal como a viralização)[326].

Dessa forma, como as plataformas não são neutras e exercem um novo tipo de controle editorial, não podem ser absolutamente isentas de responsabilidade ao tomarem decisões de moderação de conteúdo, ou mesmo decisões de personalização e segmentação de conteúdos, como sustenta Cass Sunstein. Essa nova liberdade editorial gera o dever de essas empresas privadas justificarem as razões de eventual decisão pela restrição ou priorização de determinado conteúdo, pois isso afeta a liberdade de expressão de usuários.

Como observa Tarleton Gillespie, por motivos práticos, as plataformas precisam de regras que possam ser seguidas, que façam sentido para os usuários, que forneçam à equipe de políticas um guia razoavelmente claro para decidir o que remover, que deixem espaço suficiente para conteúdo questionável que eles possam querer reter, e que possam mudar ao longo do tempo. Para o autor, isso é o que fornecerá uma justificativa satisfatória para remoções que forem contestadas, seja pelos próprios usuários ou pela sociedade em geral. Mais do que isso, articular as regras é a oportunidade mais clara para as plataformas justificarem seus esforços de moderação como legítimos[327].

Vê-se, portanto, que a discussão não é sobre se as plataformas têm ou não autonomia e liberdade editorial para moderar conteúdo. Elas têm, e definem suas políticas internas no âmbito dessas liberdades privadas. Por isso, é necessário

[325] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 123.
[326] Ibid, p. 124.
[327] GILLESPIE, Tarleton. *Custodians of the internet:* platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, 2018, p. 45.

PUC-Rio - Certificação Digital Nº 2112697/CA

pensar nos limites dessas liberdades e até onde essas empresas privadas podem ir diante de problemas e contextos específicos[328].

3.2 Autorregulação regulada: regulação estatal que deve se limitar à perspectiva procedimental da moderação de conteúdo exercida pelas plataformas de redes sociais

Como visto no Capítulo 2, a regulação da moderação de conteúdo puramente estatal e a autorregulação se mostraram insuficientes. A autorregulação foi a resposta encontrada no início da expansão da internet. Essa modalidade pretende regular a moderação de conteúdo por meio do código, que seria, para Lawrence Lessig, espécie de lei no ambiente digital[329]. A autoregulação dificulta a violação das regras do código, salvo nos casos em que seja possível *hackeá-lo*. Sem isso, na internet, por esse modelo, só será possível fazer o que o código permitir. São as próprias plataformas que criam os códigos, com base em seus interesses privados.

Com o desenvolvimento da internet, contudo, regras opacas e a ausência de transparência na atividade de moderação de conteúdo (e no desenvolvimento do "código") causaram uma crise de legitimidade das decisões tomadas pelas plataformas, que passaram a ser duramente criticada pelos usuários, exigindo uma mudança de comportamento. De fato, passou-se a perceber que a postura de total abstenção estatal sobre esse assunto não é capaz de proteger interesses públicos relevantes, como os direitos fundamentais e a democracia[330].

Por outro lado, na internet, a regulação puramente estatal implica no risco de o Estado atuar com base em seus interesses próprios, censurando discursos que entenda prejudiciais, em violação ao direito fundamental de liberdade de expressão. Tal risco se apresenta como igual ou até mesmo superior ao risco de censura privada das próprias plataformas. Luna Barroso explica que tal intervenção indevida do Estado pode se dar por diferentes meios: (i) leis excessivamente duras de responsabilidade civil dos intermediários pelo conteúdo gerado por terceiros, que

[328] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 125.
[329] LESSIG, Lawrence. *Code: version 2.0*, Basic Books, 2006, p. 122-137.
[330] KELLER, Clara Iglesias. *Regulação nacional de serviços na Internet*: exceção, legitimidade e o papel do Estado. Tese (Doutorado), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019, p. 175.

PUC-Rio - Certificação Digital Nº 2112697/CA

acabam por criar incentivos para remoção excessiva de conteúdo; (ii) leis vagas que utilizam conceitos abertos e indeterminados para impor restrições à liberdade de expressão, como "interesse público"; (iii) obrigações para que as próprias plataformas avaliem a legitimidade de pedidos de remoção de conteúdo do Estado ou de particulares; e (iv) ameaças de responsabilidade civil e penal de funcionários locais das empresas em caso de não atendimento de ordens de remoção de conteúdo[331].

De fato, a regulação puramente estatal se apresenta ineficiente e perigosa, especialmente no campo da liberdade de expressão. Afinal, a participação das plataformas de redes sociais, que desenvolvem os códigos do ciberespaço, é imprescindível para qualquer pretensão regulatória, pois os algoritmos por elas desenvolvidos podem potencializar ou até mesmo inviabilizar a efetividade de leis, já que as plataformas são "a instância final responsável por influenciar comportamentos *online*"[332]. Por isso mesmo, o debate sobre a forma de regulação deve necessariamente considerar esses atores intermediários, como esclarecem Georges Abboud e Ricardo Campos:

> Especialmente em âmbitos complexos como os das novas tecnologias, o conhecimento necessário para a tomada da decisão não se encontra no Estado, tornando assim necessária a criação de novas formas de geração do conhecimento dentro do direito regulatório estatal que incorpore o conhecimento advindo da sociedade. Em outras palavras, a constelação clássica do direito administrativo programado, por um lado, na estrutura de normas com hipótese de incidência e consequência jurídica, ou, por outro lado, num maior âmbito de ação de agências reguladoras, devem dar espaço na sociedade das plataformas para uma forma de regulação mais reflexiva, em que a observação e a incorporação de modelos de auto-organização da sociedade ganham preponderância[333].

Ainda, não há como desconsiderar que o espaço digital está em constante evolução e transformação, e regulações estatais dificilmente conseguem acompanhar isso, podendo se tornar obsoletas com facilidade. Ademais, agentes estatais geralmente não têm conhecimentos técnicos específicos acerca do funcionamento das plataformas e de seus algoritmos[334].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[331] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 220.
[332] KELLER, Clara Iglesias. *Regulação nacional de serviços na Internet:* exceção, legitimidade e o papel do Estado. Tese (Doutorado), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019, p. 158.
[333] ABBOUD, Georges, CAMPOS, Ricardo. A autorregulação regulada como modelo do Direito proceduralizado. *In:* ABBOUD, Georges, NERY NÚNIOR, Nelson. CAMPOS, Ricardo. *Fake News e Regulação*. São Paulo: Thomson Reuters Brasil, 2021, p. 143.
[334] BARROSO, Luna van Brussel. Op. cit., p. 219.

Não há como desconsiderar, por fim, que regulações estatais que objetivam proibir ou restringir excessivamente a moderação de conteúdo em redes sociais – como o Decreto Presidencial do então presidente Jair Bolsonaro, que pretendia proibir a moderação de conteúdo fora das hipóteses taxativamente previstas (Capítulo 2.5.4) –, além de poder aumentar o volume de conteúdo ilícito, propiciando ambientes tóxicos, violam o direito fundamental à livre iniciativa e à liberdade de expressão das próprias plataformas, que são empresas privadas e, como tal, devem ter a liberdade de definir qual serviço prestar e como prestar esse serviço, desde que "forneçam as informações necessárias e suficientes para que os usuários entendam, a partir de Termos de Uso claros e específicos, o que é proibido"[335]. Por isso mesmo, Clara Keller entende que "é fundamental que qualquer restrição estatal seja a mínima possível e tenha o potencial de promover o interesse público que justificou a intervenção regulatória"[336].

Nesse contexto, em que tanto a autorregulação como a regulação estatal não se mostram suficientes para lidar com os desafios da moderação de conteúdo em redes sociais, a autorregulação regulada se apresenta como uma alternativa promissora à governança da internet. O modelo da autorregulação regulada recorre à cooperação entre o Estado e os atores regulados para melhor explorar os conhecimentos dos agentes privados, ao mesmo tempo em que garante proteção a direitos fundamentais e valores de interesse público[337].

Georges Abboud e Ricardo Campos esclarecem que esse é o modelo da proceduralização, um terceiro modelo (além do direito como expressão das normas postas, com a centralidade do Estado, e do modelo da ponderação, com o direito materializado em princípios abstratos), que surge justamente "da crise do direito regulatório devido ao aumento da complexidade social"[338]:

> Especialmente em âmbitos complexos como os das novas tecnologias, o conhecimento necessário para a tomada da decisão não se encontra no Estado, tornando assim necessária a criação de novas formas de geração do conhecimento

[335] Ibid, p. 222.
[336] KELLER, Clara Iglesias. *Regulação nacional de serviços na Internet*: exceção, legitimidade e o papel do Estado. Tese (Doutorado), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019, p. 114.
[337] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 223.
[338] ABBOUD, Georges, CAMPOS, Ricardo. A autorregulação regulada como modelo do Direito proceduralizado. *In:* ABBOUD, Georges, NERY Jr., Nelson; CAMPOS, Ricardo. *Fake News e regulação*, São Paulo: Thomson Reuters Brasil, 2021, p. 142.

PUC-Rio - Certificação Digital Nº 2112697/CA

> dentro do direito regulatório estatal que incorpore o conhecimento advindo da sociedade.
> (...)
> Enquanto o modelo da ponderação incorpora em seu modelo um horizonte reduzido de formulação de novas distinções e conceitos jurídicos para orientar novas decisões, ficando a cabo de um situacionismo do caso a caso, o modelo da proceduralização foca na dimensão processual para aquisição de conhecimento para decisão em âmbitos complexos da sociedade na qual o conhecimento para decisão não decorre de simples ponderação de dois princípios abstratos[339].

Dieter Grimm pontua que a autorregulação regulada é a forma mais avançada de proceduralização, "pois promete ser mais sensível às mudanças das variáveis estruturais e explora melhor as capacidades de informação dos destinatários e a lógica inerente a cada esfera social"[340]. Com efeito, esse modelo permite que os agentes regulados com conhecimento técnico formulem, interpretem e implementem a regulação, reduzindo os custos e aumentando a eficácia regulatória[341]. Ao mesmo tempo, a autorregulação regulada endereça preocupações relacionadas com a defesa de direitos fundamentais na internet, que muitas vezes não são considerados em propostas de autorregulação pura e simples[342].

Ou seja: a participação dos agentes regulados supre a falta de conhecimento técnico do Estado, ao mesmo tempo em que a participação do Estado permite a fixação de parâmetros de interesse público[343]. Assim, o modelo da autorregulação regulada se mostra mais adequado para os fins de:

> (i) criar os incentivos adequados para que as plataformas removam conteúdo ilegal ou especialmente danoso; (ii) proteger informações e discursos legítimos, evitando qualquer intervenção estatal excessiva e modelos de responsabilização civil que levem à remoção de conteúdo em excesso; e (iii) promover a liberdade de iniciativa e a inovação, pois leis de responsabilidade civil excessivamente duras podem inviabilizar serviços lícitos que promovem o bem-estar social e impedir a ascensão de novas plataformas ou de novos serviços dentro de plataformas já existentes, por medo de responsabilização e pelos custos regulatórios[344].

PUC-Rio - Certificação Digital Nº 2112697/CA

---

339 Ibid, p. 143.
340 DIETER Grimm. Regulierte Selbstregulierung in der Tradition des Verfassungsstaates, in: *Die Verwaltung. Zeitschrift fur Verwaltungrecht und Verwaltungswissenschaften*, Caderno 4, Regulierte Selbstregulierung als Steuerungskonzept des Gewahrleistungsstaates, Duncker & Humblot, Berlim, 2001, p. 18. *In:* ABBOUD, Georges, NERY junior, Nelson; CAMPOS, Ricardo. *Fake News e regulação*, São Paulo: Thomson Reuters Brasil, 2021, p. 129.
341 BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*: o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 223.
342 BINEMBOJM, Gustavo. *Poder de polícia, ordenação, regulação*. 2. ed. Belo Horizonte: Fórum, 2017, p. 282.
343 BARROSO, Luna van Brussel. op. cit., p. 223.
344 Ibid, p. 224.

Uma forma de autorregulação regulada é regulação responsiva, desenvolvida por Ian Ayres e John Braithwaite[345]. A teoria surgiu como desenvolvimento do padrão tradicional regulatório de comando e controle entre o Estado e o administrado, creditando maior efetividade a mecanismos de autorregulação. Os autores propõem um modelo escalonado de regulações, representado por uma pirâmide. A teoria dá protagonismo a formas passivas de autorregulação, com a adoção de formas autoregulatórias com sanções na camada inferior, seguida por comandos regulatórios com sanções discricionárias e, por fim, no topo da pirâmide, o uso de sanções tradicionais não-discricionárias.

A teoria responsiva visa justamente dar voz ativa ao próprio sujeito da regulação, no caso, as plataformas de redes sociais, que deverão cumprir as regras procedimentais sobre moderação de conteúdo de forma voluntária. Isso dá autonomia às empresas, prezando pela construção de modelos de governança empresariais que exijam a internalização de custos de fiscalização, conscientização, e adequação às normas a serem impostas.

Destaca-se que a regulação estatal da moderação de conteúdo deve se concentrar não na licitude ou ilicitude do conteúdo, mas sim em procedimentos adequados para legitimar as decisões das plataformas. Esse tipo de regulação não deve pretender vedar a moderação de conteúdo feita com base nos termos de uso das plataformas, pelo contrário. É preciso preservar a autonomia dessas empresas privadas para definirem o tipo de ambiente digital que pretendem fornecer aos usuários, com base na liberdade de iniciativa e de expressão das próprias plataformas, desde que forneçam informações necessárias e suficientes aos usuários, para que entendam o que pode ou não ser feito *online*.

Para garantir que as plataformas moderem o discurso de forma sistemática e não arbitrária, a regulação estatal deve se concentrar em impor obrigações procedimentais, como a de divulgação de informações sobre a base da tomada de decisão da plataforma, bem como exigir mecanismos internos de supervisão que tornem as decisões da plataforma publicamente responsabilizáveis. Nesse sentido, Evelyn Douek defende que a regulamentação governamental deve se concentrar em legitimar os *processos* pelos quais as plataformas tomam decisões sobre o discurso, em vez de se direcionar para *substância* dessas decisões. As plataformas devem ser

[345] AYRES, Ian; BRAITHWAITE, John. *Responsive Regulation:* Transcending the Deregulation Debate. Nova Iorque: Oxford University Press, 1992.

PUC-Rio - Certificação Digital Nº 2112697/CA

exigidas a mostrar o que estão moderando de acordo com suas regras públicas, que devem ser transparentes para trazer maior *accountability* às plataformas[346].

Evelyn Douek chama isso de "*verified accountability*", cabendo às plataformas a obrigação de fazer os aspectos de sua governança transparentes e *accountable*[347]. A regulação deve, assim, focar em trazer deveres como os de transparência, devido processo legal, e isonomia para as plataformas.

De acordo com o relatório do InternetLab "armadilhas e caminhos na regulação da moderação de conteúdo", essa abordagem procedimental seria uma forma de disciplinar a atuação das plataformas quando elas agem sobre seus usuários, sem, contudo, engessá-las. "No lugar de tachar qualquer moderação como censura, essa perspectiva busca regular esse processo, reconhecendo direitos aos usuários e proporcionando mais transparência, tanto aos usuários envolvidos numa controvérsia quanto à sociedade em geral"[348].

Esse tipo de regulação com base em deveres procedimentais é contemplada, por exemplo, nos Princípios de Manila[349], documento organizado pela sociedade civil de todo o mundo. O sexto princípio do documento prevê que a "transparência e prestação de contas devem ser integradas em leis e em políticas e práticas de restrição de conteúdos", dispondo, dentre outros, que:

> e. Os intermediários devem publicar relatórios de transparência que forneçam informações específicas sobre todas as restrições de conteúdos realizadas pelo intermediário, incluindo ações tomadas devido à requisição governamental, ordens judiciais, requisições de agentes privados e a implementação de políticas de restrição de conteúdo;
> f. Nos casos em que o conteúdo tenha sido restrito em um produto ou serviço do intermediário que permita a exibição de uma notificação quando alguém tenta acessá-lo, o intermediário deve exibir uma notificação clara que explique qual conteúdo foi restrito e o motivo para tanto[350].

No mesmo sentido, entidades da sociedade civil e acadêmicos do tema publicaram, em 2018, os "Princípios de Santa Clara para transparência e prestação de contas em moderação de conteúdo"[351]. O documento sugere três tipos de

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[346] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, p. 7, available at. Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 18 out. 2022.
[347] Ibid.
[348] Disponível em: <internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 18 out. 2022.
[349] Disponível em: <https://manilaprinciples.org/pt-br.html>. Acesso em: 18 out. 2022.
[350] Ibid.
[351] Disponível em: <https://santaclaraprinciples.org/pt/cfp/>. Acesso em: 18 out. 2022.

princípios gerais a serem incorporados pelas políticas de moderação de conteúdo das redes sociais.

O primeiro se relaciona com os "números", visando a divulgação pública de estatísticas sobre a remoção de conteúdo. O documento pede que haja uma especificação mínima desse mecanismo de transparência, sugerindo que seja publicado, por exemplo, o número de postagens reportadas (*flagged*) e suas autorias (se por governos, usuários, entidades credenciadas, etc.) em relatórios trimestrais.

Nesse ponto, vale destacar que Tarleton Gillespie também defende que a moderação de conteúdo praticada pelas redes sociais seja mais transparente, o que, para o autor, não significa apenas a falta de opacidade. A transparência exige a criação de novas ferramentas para tornar as informações processuais visíveis, mas discretas:

> As plataformas devem fazer um compromisso radical de devolver os dados que eles já têm para mim em uma forma legível e acionável, tudo o que elas poderiam me dizer contextualmente sobre porque um post está lá e como devo avaliá-lo. Nós já pagamos por essa transparência, com nossos dados[352].

O segundo se relaciona com "notificação", ou seja, sugere que usuários que tiveram postagens derrubadas ou contas suspensas devem ser notificados sobre os motivos de tais decisões. O documento prevê que tais notificações contenham, inclusive, identificação precisa da URL, a cláusula sobre moderação de conteúdo supostamente violada pelo usuário, e como a plataforma chegou a tal decisão (se por processo automático, se por ordem judicial, se por reclamação de outro usuário etc.), e o processo para recorrer dessa decisão.

O fornecimento dessas razões é extremamente relevante para que o Poder Judiciário possa avaliar a razoabilidade da decisão tomada pelas plataformas. Como visto, as plataformas de redes sociais são empresas privadas que têm autonomia editorial para estabelecer suas próprias regras sobre moderação de conteúdo, que podem ser, eventualmente, mais restritivas do que os padrões de direito público. Por isso mesmo, pode e deve o Poder Judiciário se pronunciar sobre a razoabilidade de eventual restrição de conteúdo, e para que isso se dê da melhor forma, é preciso que as decisões das plataformas sejam motivadas[353].

---

[352] GILLESPIE, Tarleton. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 199.

[353] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 172.

PUC-Rio - Certificação Digital Nº 2112697/CA

Por fim, o terceiro princípio geral se relaciona com "apelação", para que as plataformas promovam uma efetiva oportunidade de recurso aos usuários, em tempo razoável nos casos de derrubada de postagem ou suspensão de perfil. Tal processo de revisão da decisão deve incluir a revisão humana por pessoas não envolvidas na decisão inicial (especialmente em casos em que esta decisão é feita de forma automatizada), e notificar o usuário sobre o resultado desse processo.

Existe uma correlação entre o amadurecimento de demandas da sociedade civil e especialistas no sentido de um devido processo digital e medidas concretas já tomadas nesse sentido pelas plataformas[354]. O Facebook, por exemplo, pouco antes da publicação dos Princípios de Santa Clara, publicou suas diretrizes e parte de seus critérios de moderação de conteúdo, até então restritos a seus ambientes internos, além de ampliar os processos de recursos contra suas decisões.

No âmbito transnacional do direito das plataformas, os mecanismos de devido processo legal têm sido, principalmente, resultado de iniciativas de autolimitação, transparência e prestação de contas pelas próprias empresas, na tentativa de dar mais legitimidade às suas decisões[355].

A melhor ilustração dessas iniciativas é a criação, pelo Facebook, de um órgão independente, formado por pessoas de fora da empresa, para tomar decisões finais, transparentes, e vinculantes sobre moderação de conteúdo. O órgão (*Oversight Board*) é tratado em detalhes no Capítulo 3.3. Antecipa-se, desde já, que o maior benefício do órgão foi garantir um foro público para apresentações de razões e argumentos que sustentem publicamente determinada decisão, aumentando assim a possibilidade de controle sobre as regras e decisões da plataforma sobre moderação de conteúdo.

PUC-Rio - Certificação Digital Nº 2112697/CA

3.3 A necessária criação de órgãos de fiscalização independentes

A regulação estatal procedimental se apresenta como a melhor opção diante dos desafios da moderação de conteúdo, mas também traz dificuldades. Não existe modelo regulatório perfeito. Uma regulação muito abrangente acabará alcançando conteúdos perfeitamente legítimos e lícitos. Uma regulação menos abrangente

[354] Ibid, p. 142.
[355] Ibid.

permitirá maior proliferação de conteúdos tóxicos e ilícitos. Além disso, como visto no Capítulo 1.4, a forma de moderação de conteúdo automatizada, embora necessária diante do desafio de escala, leva à remoção de conteúdo muitas vezes legítimo, além de conter os vieses dos algoritmos, aumentando a margem de erro das decisões. Esses erros temporários, contudo, podem ser inevitáveis e até aceitáveis para que se tenha do outro lado decisões tomadas de forma rápida[356].

A moderação de conteúdo puramente humana, por outro lado, embora seja capaz de minimizar os erros da moderação automatizada, se mostra impossível, na prática, diante do volume de conteúdo postado diariamente nas plataformas. A criação de um sistema que permita recursos com oportunidades de apresentar informações adicionais em todo e qualquer caso, por exemplo, não atenderia ao propósito de maior devido processo legal. Poderia haver mais inconsistência nas decisões, demora na revisão, e o nível de explicação fornecido seria menor. Isso é normal em qualquer sistema: certeza em um caso isolado ou maior certeza no sistema em geral. Por isso, para Evelyn Douek, o que significa "devido processo legal" deve ser determinado contextualmente[357].

Não se pode esquecer, ainda, que uma regulação perfeita sobre liberdade de expressão é inviável, pois nos casos mais controvertidos haverá divergência razoável sobre a licitude ou ilicitude de determinado conteúdo[358]. Com efeito, parâmetros substantivos de direitos fundamentais estão sempre em aberto, pois não são definidos de antemão[359]. Tal divergência sobre a legitimidade e licitude de um conteúdo traz o risco de que governos se aproveitem dessas divergências para promover interesses antidemocráticos. Ainda, a possibilidade de fiscalização pelo Poder Judiciário traz riscos de decisões contraditórias e sem capacidade institucional para analisar as restrições técnicas do setor[360].

Diante desses pontos, a regulação da moderação de conteúdo deve ter como objetivo encontrar um modelo que seja capaz de otimizar a ponderação entre

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[356] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, p. 10, available at. Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation>. Acesso em: 18 out. 2022.

[357] Ibid.

[358] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 228.

[359] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 140.

[360] BARROSO, Luna van Brussel. op. cit., p. 283.

direitos fundamentais de usuários e plataformas, reconhecendo que sempre haverá casos de erro ou casos em que não será possível alcançar um consenso[361].

Por isso, diante dos desafios enfrentados, torna-se importante que as decisões das plataformas sobre moderação de conteúdo sejam fiscalizadas por órgãos independentes, de modo a preservar a liberdade de expressão diante da centralidade das plataformas na esfera pública. Tais órgãos devem servir como fiscalizadores do funcionamento do sistema de deveres procedimentais proposto nesse estudo. Assim, deverão fiscalizar se as plataformas estão cumprindo com os requisitos de transparência, devido processo legal e isonomia, e os deveres mínimos para combater ou minimizar os impactos de conteúdo danoso[362]. Nesse sentido, assim sustenta relatório do InternetLab:

> Qualquer que seja o caminho adotado, é crucial que a fiscalização das plataformas seja realizada por órgãos com independência. Isso é fundamental para preservação da liberdade de expressão, considerando a centralidade das plataformas na esfera pública, quando pensamos numa chave de defesa de um ambiente aberto e democrático na internet[363].

A criação de órgãos de supervisão das decisões da plataforma se mostra útil pois pode fornecer um fórum para explicação de erros e escolhas entre os *trade-offs* envolvidos em qualquer projeto de sistema. Por exemplo, tomar decisões rápidas é necessário para lidar com o volume, mas tem como *trade off* aumentar as chances de erro. Dar possibilidade de recurso em todos os casos, por sua vez, aumenta as chances de uma decisão justa e acertada, mas também aumenta o tempo de resolução do conflito, e pode incentivar regras de moderação de conteúdo mais frouxas.

Um órgão fiscalizador independente pode, ainda, fazer com que regras abstratas sejam mais compreensíveis por meio de decisões disponibilizadas em casos concretos, como uma espécie de jurisprudência das decisões do órgão. Mas, quando se fala em "jurisprudência", cumpre lembrar que um único órgão não é capaz de promover transparência em todos os casos, não só pelo desafio da escala, mas também porque há casos que devem ser sigilosos por preocupações de privacidade, por exemplo.

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[361] DOUEK, Evelyn. *Governing online speech*. Columbia Law Review, v. 121, n. 03, 2021. Disponível em: <https://columbialawreview.org/wp-content/uploads/2021/04/Douek-Governing_Online_Speech-from_Posts_As-Trumps_To_Proportionality_And_Probability.pdf>. Acesso em 24 nov. 2022.

[362] Ibid.

[363] Disponível em: <internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 18 out. 2022.

De toda forma, é preciso haver um mecanismo que garanta que, mesmo nesses casos, as regras da plataforma estão sendo aplicadas de forma consistente e imparcial. Um órgão de fiscalização independente pode ser esse garantidor[364]. Com efeito, uma verificação no estilo "judicial" das decisões da plataforma pode melhorar suas políticas e ajudar a remediar a falta de confiança criada pelo histórico de opacidade e ofuscação na moderação de conteúdo até o momento. O órgão fará as decisões da plataforma terem suas justificativas publicizadas, o que trará mais legitimidade para as decisões.

Um órgão de fiscalização também pode dar maior visibilidade às normas da comunidade e difundir a pressão pública (mitigando o problema de uma plataforma privada ser o "árbitro da verdade").

É preciso apontar que o órgão de fiscalização não conseguirá resolver a questão dos diferentes contextos, criando regras globais sobre o conteúdo a ser moderado. Esses órgãos, contudo, não devem se concentrar na harmonização de regras entre distintas comunidades, mas em trazer perspectivas diversas sobre como as regras da plataforma precisam ser aplicadas de forma diferente em contextos diferentes[365].

Foi nesse sentido que a Meta criou, em 2018, o Comitê Supervisor ("*Oversight Board*"). A ideia da plataforma de estabelecer uma "corte independente" em seu sistema de governança privada remete a um sistema de separação de poderes, no qual um órgão julgador independente supervisiona as demais funções, como a legislativa e a executiva[366]:

> À medida que sua comunidade cresceu para mais de 2 bilhões de pessoas, ficou cada vez mais claro para a empresa Facebook que não deveria tomar tantas decisões sobre fala e segurança online por conta própria. O Comitê de Supervisão foi criado para ajudar o Facebook a responder a algumas das perguntas mais difíceis sobre a liberdade de expressão online: o que remover, o que deixar e por quê.
> O comitê usa seu julgamento independente para apoiar o direito das pessoas à liberdade de expressão e garantir que esses direitos sejam devidamente respeitados. As decisões do comitê de manter ou reverter as decisões de conteúdo do Facebook

[364] DOUEK, Evelyn. *Verified Accountability:* Self-regulation of content moderation as an answer to the special problems of speech regulation. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, p. 12, available at Disponível em: <https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulat>. Acesso em: 18 out. 2022.
[365] Ibid.
[366] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 143.

PUC-Rio - Certificação Digital Nº 2112697/CA

> serão obrigatórias, o que significa que o Facebook terá que implementá-las, a menos que isso viole a lei[367].

O processo que levou à criação desse Comitê envolveu consultas a diversos especialistas. Em seu relatório de junho de 2019, que sintetizou a primeira etapa dessa consulta, a Meta resumiu o *feedback* que recebeu de seis *workshops* aprofundados e 22 mesas redondas, com a participação de mais de 650 pessoas de 88 países diferentes. Ao longo desse período, a Meta também teve discussões pessoais com mais de 250 pessoas e coletou mais de 1.200 contribuições sobre a formação do Comitê. Desde esse relatório, a Meta continuou a consultar especialistas sobre tópicos como a confiança do Comitê, estatutos e membros[368].

O Comitê, quando completamente formado, terá 40 membros de todo mundo, de modo a representar um conjunto diversificado de disciplinas e experiências. Esses membros terão o poder de selecionar casos para revisão e manter ou reverter as decisões de moderação de conteúdo da Meta[369].

Para assegurar a independência do Comitê em sua tomada de decisão, tanto ele quanto sua administração são financiados por um fundo independente e apoiados por uma empresa independente e separada da empresa. A Meta montou um fundo independente de cerca de 130 milhões de dólares que irá financiar suas atividades pelos próximos 6 anos, incluindo uma equipe própria de funcionários[370].

O Comitê tem autoridade para decidir se o Facebook e o Instagram devem permitir ou remover determinado conteúdo. Essas decisões são vinculantes, a menos que sua implementação possa violar a lei. O Comitê também pode optar por emitir recomendações sobre as políticas de conteúdo da empresa, que, embora não vinculantes, devem ser observadas pelo Facebook ou Instagram no prazo de 30 dias, mediante a indicação fundamentada de como foram consideradas internamente e de quais não serão implementadas.

Os indivíduos poderão recorrer das decisões sobre moderação de conteúdo do Facebook e do Instagram ao Comitê. Qualquer pessoa cujo conteúdo seja selecionado para revisão pelo Comitê terá a oportunidade de compartilhar uma declaração explicando sua posição. Ainda, o Comitê deve compartilhar

[367] Disponível em: <https://www.oversightboard.com/>. Acesso em: 18 out. 2022.
[368] Disponível em: <https://www.oversightboard.com/>. Acesso em: 18 out. 2022..
[369] Ibid.
[370] Why Mark Zuckerberg's Oversight Board may kill his political and policy", reportagem publicada por Wired, em 28/01/2020.

PUC-Rio - Certificação Digital Nº 2112697/CA

publicamente declarações escritas sobre suas decisões e justificativas. Cada decisão será publicada e arquivada no *site* do Comitê. Além disso, o Comitê divulgará relatórios anuais sobre seu trabalho[371].

Recursos podem ser submetidos ao Comitê por usuários ou pela própria Meta, e o Comitê tem liberdade para decidir quantos e quais casos serão analisados, já que a empresa recebe cerca de um milhão de denúncias aos termos de uso por dia[372]. As escolhas serão feitas por um comitê de seleção, composto de modo rotativo por uma parcela dos integrantes do Comitê. O objetivo é que o Comitê se dedique aos casos mais relevantes e paradigmáticos. O regimento interno já publicado prevê prazos de julgamento, que podem ser abreviados quando houver urgência a pedido da empresa.

As decisões do Comitê serão vinculantes ao respectivo caso concreto. Apenas não serão acatadas decisões em casos em que isso implique na violação de alguma legislação pela Meta.

Já existem hoje (novembro de 2022) 33 decisões do Comitê publicadas. Dessas, apenas 9 foram no sentido de manter a decisão da Meta[373]. A mais recente decisão publicada é relacionada à decisão da Meta de restaurar uma postagem do Instagram contendo o vídeo de uma mulher sendo agredida sexualmente por um grupo de homens.

Em março de 2022, uma conta do Instagram que se descreve como uma plataforma para perspectivas *dalits* postou um vídeo da Índia mostrando uma mulher sendo agredida por um grupo de homens. Os *dalits* já foram chamados de "intocáveis" e enfrentaram a opressão sob o sistema de castas. O rosto da mulher não é visível no vídeo e não há nudez. O texto que acompanha o vídeo afirma que uma "mulher tribal" foi abusada sexualmente em público e que o vídeo se tornou viral. Depois que um usuário denunciou a postagem, a Meta a removeu por violar a política de exploração sexual adulta, que proíbe conteúdo que "retrate, ameace ou promova violência sexual, agressão sexual ou exploração sexual"[374].

PUC-Rio - Certificação Digital Nº 2112697/CA

[371] Disponível em: <https://www.oversightboard.com/>. Acesso em: 18 out. 2022.
[372] DOUEK, Evelyn. *Facebook's Oversight Board:* move fast with stable infrastructure and humility, Noth Carolina Journal of Law and Technology Volume 21 (2019), p. 11.
[373] Disponível em: <https://www.oversightboard.com/decision/>. Acesso em: 20 nov. 2022.
[374] Disponível em: <https://www.oversightboard.com/decision/IG-KFLY3526/>. Acesso em: 20 nov. 2022.

Um funcionário da Meta sinalizou a remoção do conteúdo por meio de um canal de denúncias interno ao saber disso no Instagram. As equipes internas da Meta revisaram o conteúdo e aplicaram um "subsídio de noticiabilidade", que permite que o conteúdo da violação permaneça nas plataformas da Meta se for relevante e de interesse público. A Meta então restaurou o conteúdo, colocando o vídeo atrás de uma tela de aviso que impede que menores de 18 anos o vejam, e posteriormente encaminhou o caso ao Comitê.

O Comitê entendeu que restaurar o conteúdo da plataforma, com a tela de aviso, é consistente com os valores e as responsabilidades de direitos humanos da Meta. O Comitê reconheceu que a rede social é um meio importante de documentar tal violência e discriminação e que o conteúdo neste caso pareceu ter sido postado para aumentar a conscientização. Assim, decidiu por manter a decisão da Meta[375].

A segunda decisão mais recente do Comitê foi pela revogação da decisão da Meta de remover um vídeo no Instagram mostrando as consequências de um ataque terrorista na Nigéria. Em 5 de junho de 2022, um usuário do Instagram na Nigéria postou um vídeo mostrando corpos imóveis e ensanguentados no chão, como resultado de um ataque terrorista a uma igreja no sudoeste da Nigéria, no qual pelo menos 40 pessoas foram mortas e muitas outras ficaram feridas. O conteúdo foi postado no mesmo dia do ataque. Os comentários no *post* incluíram orações e declarações sobre segurança na Nigéria.

Os sistemas automatizados da Meta revisaram o conteúdo e aplicaram uma tela de aviso. No entanto, o usuário não foi alertado, pois usuários do Instagram não recebem notificações quando telas de aviso são aplicadas. Posteriormente, o usuário adicionou uma legenda ao vídeo. Descreveu o incidente como "triste" e usou várias *hashtags*, incluindo referências a colecionadores de armas de fogo, alusões ao som de tiros e ao jogo *live-action "airsoft"* (onde as equipes competem com armas simuladas).

Pouco depois, um dos bancos do *Media Matching Service* da Meta, um "banco de escalações", identificou o vídeo e o removeu. Os bancos do *Media Matching Service* podem combinar automaticamente as postagens dos usuários com o conteúdo que foi considerado violador anteriormente. Se o conteúdo de um "banco de escalações" foi considerado violado pelas equipes internas especializadas

[375] Ibid.

PUC-Rio - Certificação Digital Nº 2112697/CA

da Meta, qualquer conteúdo correspondente é identificado e imediatamente removido. O usuário apelou da decisão para a Meta e um revisor humano confirmou a remoção. O usuário então recorreu ao Comitê.

Quando o Comitê aceitou o caso, a Meta revisou o conteúdo no "banco de escalações", concluiu que não havia violação, e o removeu do banco. Ainda assim, manteve a decisão de remover o vídeo nesse caso, pois entendeu que as *hashtags* de sua postagem poderiam ser lidas como "glorificando a violência e minimizando o sofrimento das vítimas", o que viola várias políticas, como a de conteúdo violento e gráfico, que proíbe comentários sádicos.

O Comitê, contudo, decidiu pela restauração da postagem com uma tela de aviso de "conteúdo perturbador". Considerou que a restauração do vídeo é consistente com as Diretrizes de Comunidade da Meta, seus valores, e responsabilidades com direitos humanos. Apontou que a Nigéria está passando por uma série contínua de ataques terroristas e que o governo nigeriano suprimiu a cobertura de alguns deles. O Comitê concluiu, portanto, que nesse contexto a liberdade de expressão é particularmente importante.

O Comitê também recomendou que a Meta: (i) revise a linguagem na política pública de conteúdo violento e explícito para garantir que esteja alinhada com a orientação interna para moderadores; e (ii) notifique os usuários do Instagram quando uma tela de aviso for aplicada ao seu conteúdo e forneça a justificativa da política específica para isso[376].

Ao avaliar a criação do Comitê Supervisor da Meta, Evelyn Douek aponta dois benefícios:

> Primeiro, pode ajudar a iluminar defeitos na formulação de políticas pelo Facebook, removendo bloqueios (como pontos-cego ou inércia) no 'processo legislativo' que leva à formulação de 'padrões de comunidade' Em segundo lugar, ao possibilitar um fórum independente para discussões de decisões polêmicas sobre moderação de conteúdo, o Comitê Supervisor pode ser um importante fórum para o processo de razão pública necessário para que as pessoas em uma comunidade plural aceitem as regras que as governam, mesmo se elas discordem da substância dessas regras[377].

No Brasil, autores como Carlos Ari Sundfeld e Luna Barroso defendem que o CGI.br poderia desempenhar a função de um órgão independente fiscalizador do

[376] Disponível em; <https://www.oversightboard.com/decision/IG-KFLY3526/>. Acesso em: 20 nov. 2022.
[377] DOUEK, Evelyn. *Facebook's Oversight Board:* move fast with stable infrastructure and humility, Noth Carolina Journal of Law and Technology Volume 21 (2019), p. 7.

PUC-Rio - Certificação Digital Nº 2112697/CA

sistema da autorregulação regulada. Com efeito, entende-se que o Poder Judiciário não tem as melhores condições para atuar como fiscalizador desse novo modelo de regulação, pois apresenta o risco de decisões conflitantes e não a tem capacidade institucional para analisar as restrições técnicas do setor. A fiscalização deve ser atribuída a um órgão especializado, com representação majoritária da sociedade civil.

O objetivo desse órgão fiscalizador não deve ser responsabilizar plataformas por eventuais violações pontuais e específicas, mas realizar uma análise sistêmica do modelo instituído e administrado pelas plataformas e auditar ou avaliar relatórios de transparência[378]. É importante, ainda, que o órgão tenha flexibilidade regulatória para permitir a adaptação de suas recomendações, diretrizes e decisões, de acordo com o avanço tecnológico e social.

Por isso, considera-se que o CGI.br tem capacidade de atuar como órgão supervisor do sistema de autorregulação regulada da moderação de conteúdo exercida pelas plataformas de redes sociais no Brasil. Trata-se de órgão já existente, o que reduziria os custos para adaptação da estrutura, sendo uma organização sem personalidade jurídica responsável por coordenar e atribuir endereços IP em território nacional e por coordenar e registrar nomes de domínio usando o ".br". É, ainda, formado por membros do governo, em número minoritário, e por representantes do setor empresarial, acadêmico, terceiro setor e comunidade científica e tecnológica, tratando-se, portanto, de órgão especializado, independente, e reconhecido internacionalmente.

Carlos Ari Sundfeld defende o modelo como sendo não estatal de corregulação, pluriparticipativo e consensual, que permite uma gestão compartilhada dos rumos do sistema da internet no Brasil[379].

Essa proposta foi, inclusive, incluída no substitutivo apresentado pelo Grupo de Trabalho para Aperfeiçoamento da Legislação Brasileira de Internet, no Congresso Nacional. Nessa proposta, o professor Marcos Dantas Loureiro defende que seja criado dentro do CGI um Conselho de transparência e responsabilidade e

[378] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital:* o impacto das mídias sociais no mundo contemporâneo. Belo Horizonte: Fórum, 2022, p. 283.
[379] SUNDFELD, Carlos Ari; ROSILHO, André. *A governança não estatal da internet e o direito brasileiro*. *RDA*, Rio de Janeiro, v. 270, p. 41/79, set./dez. 2015.

PUC-Rio - Certificação Digital Nº 2112697/CA

um centro de estudo e pesquisa específicos para o exercício das funções de regulação das plataformas digitais[380].

3.4. Responsabilidade civil pelo exercício da moderação de conteúdo condicionada à não observância das regras procedimentais

Como visto no Capítulo 1.3.2, o Marco Civil da Internet, em seu artigo 19, isenta as plataformas de rede social da responsabilidade civil por conteúdo gerado por terceiro, salvo em caso de descumprimento de ordem judicial específica. Ao criar um regime de responsabilidade civil favorável aos intermediários, legislações como a brasileira colaboram com um regime de governança privada autônoma, já que diminuem os custos e riscos legais de suas decisões.

Embora não tenha sido expresso nesse sentido, o Marco Civil da Internet não proíbe que plataformas de redes sociais moderem conteúdo de terceiros[381]. E nem poderia, eis que, como visto no Capítulo 1.1, tal atividade é fundamental para garantir um espaço digital saudável e a própria liberdade de expressão digital.

De todo modo, se uma plataforma decide moderar conteúdo dentro de seu campo de autonomia, assume o risco jurídico da avaliação da licitude ou constitucionalidade de seu comportamento por parte do Poder Judiciário[382].

Por isso mesmo, no julgamento da apelação cível nº 0000447-46.2016.8.24.0175, o TJSC[383], em caso envolvendo a derrubada de uma sátira musical pelo Youtube, afastou a tese do Google de que o Marco Civil da Internet traz a possibilidade de responsabilidade civil apenas em casos de descumprimento de ordens judiciais, como se as decisões autônomas de derrubada da empresa fossem cobertas por uma espécie de imunidade desse tipo de responsabilização.

PUC-Rio - Certificação Digital Nº 2112697/CA

---

[380] Aperfeiçoamento Legislação Brasileira - Internet - *Fiscalização da lei:* quem regula? Disponível em: <https://www.youtube.com/watch?v=fO0PUgpFPUM>. Acesso em: 10 nov. 2022.

[381] "Pode-se afirmar, portanto, que no artigo 19 do MCI: (...) a remoção de conteúdo não dependerá exclusivamente de ordem judicial, de forma que o provedor poderá, a qualquer momento, optar por retirar o conteúdo caso ele seja contrário aos termos de uso de sua plataforma." SOUZA, Carlos Affonso Souza; TEFFÉ, Chiara Spadaccini de. *Responsabilidade dos provedores por conteúdos de terceiros na internet*. Disponível em: <https://www.conjur.com.br/2017-jan-23/responsabilidade-provedor-conteudo-terceiro-internet>. Acesso em: 8 Jan. 2022.

[382] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais*. O problema jurídico da remoção de conteúdo pelas plataformas. Belo Horizonte: Editora Dialética, 2021, p. 170.

[383] BRASIL. TJSC. Apelação Cível nº 000047-46.2016.8.24.0175. 3ª Câmara de Direito Civil. j. em 06/02/2018.

Não é esse o caso. Embora o Marco Civil da Internet não proíba a atividade de moderação de conteúdo das redes sociais, as plataformas devem observar os deveres procedimentais a serem impostos por lei. Caso não observem tais deveres, poderão ser responsabilizadas por essa atividade.

Por essa razão, os fundamentos dados pelas plataformas para restrição de uma publicação importam para que o Poder Judiciário possa avaliar a razoabilidade da medida. Nesse sentido, uma sentença de primeiro grau do TJMG condenou o Facebook a reativar o perfil de um escritor e a indenizá-lo em R$ 5 mil por danos morais pelo período que sua página ficou fora do ar. A decisão foi fundamentada principalmente com base no fato de que o Facebook não havia informado quais teriam sido as regras de seus termos de uso violadas, nem ao usuário, nem ao Judiciário, durante sua defesa processual[384].

Como a moderação de conteúdo é uma atividade necessária para preservar a liberdade de expressão *online* e não é proibida pelo Marco Civil da Internet, plataformas não podem ser responsabilizadas civilmente pela moderação de conteúdo feita com base nas regras internas. Caberá ao Poder Judiciário, contudo, avaliar se a medida obedeceu as regras procedimentais a serem impostas por lei, como a devida transparência e o direito de resposta do usuário. Apenas em casos de descumprimento das regras procedimentais que as plataformas poderão ser responsabilizadas pela moderação de conteúdo feita[385].

Nesse mesmo sentido, Juliano Maranhão, Juliana Abrusio e Ricardo Campos observam que as plataformas não podem ser responsabilizadas por violações de terceiros, que sempre irão ocorrer. Isso porque a "responsabilidade não pode ser pelo resultado da moderação de conteúdo, mas apenas procedimental"[386].

PUC-Rio - Certificação Digital Nº 2112697/CA

[384] BRASIL. TJMG. Processo nº 5000133-18.2019.8.13.0045, 2º Juizado Especial Cível da Comarca de Caeté. Sentença proferida em 17/01/2020.

[385] "Por outro lado, cada vez mais se reconhece que essa curadoria de conteúdo implica responsabilidade das plataformas pela adoção dos procedimentos adequados de governança e meios técnicos disponíveis para evitar abusos e mitigar impactos provocados por conteúdo nocivo propagado". CAMPOS, Ricardo. et al. *Regulação de "fake news" no Brasil*. Disponível em <https://institutolgpd.com/wp-content/uploads/2021/10/Regulacao-de-FAKE-NEWS-no-Brasil.pdf. Acesso em 15 nov. 2022.

[386] MARANHÃO, Juliano. et al. *Atribuição de responsabilidade das plataformas no combate às Fake News*. Disponível em: <https://zelaconsulting.com/wp-content/uploads/2021/08/ABRUSIO_Atribuicao-de-responsabilidade-das-plataformas-no-combate-as-fake-news.pdf>. Acesso em: 10 nov. 2022.

CONSIDERAÇÕES FINAIS

1. O objetivo do presente trabalho é analisar a regulação da moderação de conteúdo praticada pelas plataformas de redes sociais, de forma que promova a liberdade de expressão. Trata-se de discussão relevante uma vez que a internet oferece oportunidade inédita de realização de maximização dos benefícios sociais do exercício desse direito fundamental. De fato, as redes sociais permitiram que qualquer indivíduo tenha voz ativa perante todo o mundo, em uma forma de autocomunicação de massa. Por isso, a internet e o advento das redes sociais contribuíram positivamente para o pluralismo político e para manifestações sociais e políticas ao redor do mundo. Elas, contudo, também potencializam os danos que discursos abusivos causam aos indivíduos e às instituições. Dessa forma, é preciso pensar em modelos de regulação que fomentem os aspectos positivos das redes sociais, e minimizem os aspectos negativos.

2. A concepção tradicional da liberdade de expressão foi pensada para um mundo em que a informação era escassa e a participação no debate público dependia de investimentos financeiros elevados e de disputas de meios escassos, como frequências de rádio. Nessa concepção, predominava o entendimento de que a intervenção do Estado sobre o discurso seria uma ameaça à autonomia e à democracia. A liberdade era vista, assim, como uma liberdade negativa, que impunha ao Estado um dever de abstenção, sob o fundamento de que atribuir a instituições políticas o poder para decidir o que pode ou não ser dito é perigoso, arbitrário e ilegítimo.

3. As redes sociais e o advento de novas tecnologias acabaram por encerrar a dependência havida nos veículos de mídias tradicionais. As redes sociais se tornaram as novas praças públicas, possibilitando a criação de um espaço para debate público de forma *online*, sem controle editorial prévio. O discurso saiu, assim, da centralização das grandes mídias tradicionais, para descentralização do espaço digital, facilitando sua democratização, dando voz a minorias e diversificando o debate público.

4. Por outro lado, o crescimento das redes sociais e seu uso por bilhares de pessoas ao redor do mundo também permitiu a apropriação dessas comunidades abertas para uso abusivo, com ampla disseminação de discursos de ódio, notícias falsas, e conteúdo ilegal. Tornou-se necessário, assim, que as plataformas de redes

PUC-Rio - Certificação Digital Nº 2112697/CA

sociais controlassem, de alguma maneira, o conteúdo nelas postado. Começaram, então, a impor termos e condições de uso para definir os valores e normas de cada plataforma, moderando o conteúdo postado por terceiros. Com o tempo, as plataformas de redes sociais, que no início sugiram como meras empresas de tecnologia, passaram a exercer controle sobre o discurso, tornando-se, para alguns autores, verdadeiras governantes de espaços digitais.

5. A então descentralização sobre o discurso esperada com o início da internet acabou por ser novamente centralizada, dessa vez nas mãos de poucas empresas de tecnologia. Essa nova posição de controle das plataformas de redes sociais sobre o discurso público trouxe novos contornos ao debate sobre a liberdade de expressão. Se antes havia um controle dual, exercido pelos Estados e pelas mídias tradicionais, o controle passou a ser triangular. Governos e Estados seguem em uma ponta do controle, enquanto em outra permanecem os oradores. Estes, contudo, não se limitam mais aos veículos de mídias tradicionais. Há hoje uma terceira ponta, na qual se encontram as plataformas de redes sociais, provedores intermediários que fornecem a infraestrutura que permite publicações de conteúdo por usuários.

6. Essa nova dinâmica triangular trouxe novos desafios para liberdade de expressão. Como visto, se a internet, em sua origem, foi pensada como um espaço em que as pessoas podiam publicar conteúdo de forma livre, isso foi gradativamente alterado, na medida em que as plataformas passaram a exercer uma espécie de controle sobre esse conteúdo com base em suas regras internas, por meio da moderação de conteúdo, que é necessária para assegurar a liberdade de expressão. As plataformas devem moderar conteúdo, seja para proteger um usuário de outro, ou um grupo de seu antagonista, seja para remover conteúdo ofensivo ou ilegal, ou seja para apresentar sua melhor forma para novos usuários e para o público em geral. Assim, é preciso pensar em um novo modelo de regulação dessa atividade, que tem impactos diretos na liberdade de expressão.

7. A regulação da moderação de conteúdo pode ser feita com base em três principais abordagens: (i) leis antitruste, que buscam promover a concorrência e criar os incentivos mercadológicos para que as plataformas digitais ajam de forma alinhada aos interesses de usuários; (ii) leis de proteção à privacidade, que garantem aos usuários maior controle sobre seus dados e/ou limitam o potencial de direcionamento de conteúdo pelas plataformas; e (iii) leis sobre responsabilização

PUC-Rio - Certificação Digital Nº 2112697/CA

de intermediários pelo conteúdo postado por terceiros, que pretendem criar um modelo de responsabilização civil que crie os incentivos adequados para que as plataformas promovam a liberdade de expressão, ao mesmo tempo em que combatam conteúdo danoso.

8. Sem desconsiderar a importância das duas primeiras abordagens, o objeto do presente trabalho está restrito ao terceiro plano. Assim, importa saber: o Marco Civil da Internet permite que redes sociais removam conteúdo por decisão própria? Se sim, como regular a moderação de conteúdo pelas redes sociais? Quais os parâmetros normativos que devem nortear as condutas das redes sociais à luz da liberdade de expressão? Podem essas empresas ser responsabilizadas por, ao aplicarem seus termos de uso, removerem determinado conteúdo das redes, que não seja ilegal?

9. O Marco Civil da Internet é a principal lei sobre uso da internet no Brasil, e tem compromisso explícito com a preservação da liberdade de expressão. O regime de responsabilidade civil das plataformas de redes sociais por conteúdo gerado por terceiro fixado pelo Marco Civil da Internet é o da responsabilidade subjetiva dos provedores de aplicações de internet por conteúdo de terceiros. Estes apenas serão responsabilizados se, notificados judicialmente de um conteúdo ilícito, não o remover em prazo determinado, ou seja, por um ato omissivo, em caso de inércia após receber decisão judicial específica.

10. A obrigatoriedade de remoção do conteúdo apenas após ordem judicial estimula, em parte, que os provedores de aplicações não removam o material – que esteja em conformidade com seus termos de uso e políticas internet – apenas porque o mesmo gerou uma notificação e incentiva, assim, que a vítima busque o Poder Judiciário e fundamente os motivos pelos quais um determinado conteúdo precisa ser removido. Do contrário, os próprios provedores de aplicação estariam autorizados a decidir se um conteúdo impugnado, que não viola as políticas internas da plataforma, causa ou não um dano e se pode ou não ser exibido, o que por certo contaria com critérios muito subjetivos, a prejudicar a diversidade e o grau de inovação na internet, e podendo constituir censura privada. Além disso, poderia implicar sério entrave para o desenvolvimento de novas alternativas de exploração e comunicação, as quais poderiam não ser desenvolvidas em razão do receio de futuras ações compensatórias.

PUC-Rio - Certificação Digital Nº 2112697/CA

11. Embora não tenha sido explícito, o Marco Civil da Internet não contém nenhuma proibição para que as plataformas digitais possam remover ou moderar conteúdo com base em suas regras internas aceitas pelos usuários ao ingressarem nas respectivas plataformas. Entende-se, portanto, que plataformas digitais podem (e devem, para permitir um ambiente online saudável) moderar o conteúdo postado pelos usuários com base em seus termos de uso e diretrizes de comunidade.

12. Existem diferentes formas de moderação de conteúdo utilizadas pelas redes sociais, que caracterizam a regulação e o controle do discurso público exercidos por essas empresas privadas. Essas formas representam as possibilidades e os desafios enfrentados pelas plataformas ao lidar com uma escala massiva e sem precedentes de conteúdo de diferentes contextos em seus ambientes.

13. As plataformas de redes sociais enfrentam atualmente desafios de escala na moderação de conteúdo. Trata-se do desafio de moderar conteúdo gerado por bilhões de usuários espalhados pelo globo, em contextos culturais, sociais e econômicos distintos. Com efeito, as redes sociais atuam de forma global e têm de lidar com um volume de discurso sem precedentes, o que torna a revisão humana do conteúdo muitas vezes difícil de ser implementada em termos práticos. Para enfrentar o desafio de escala, tornou-se necessária a criação de ferramentas automatizadas de inteligência artificial, que são atualmente determinantes na moderação de conteúdo. Tais ferramentas, contudo, são naturalmente passíveis de erros e não têm a mesma precisão na capacidade de interpretar o contexto de uma publicação como teriam revisores humanos, por exemplo.

14. As plataformas de redes sociais enfrentam, ainda, desafios relacionados ao baixo *accountability* e à transparência insuficiente na tomada de suas decisões. As plataformas de redes sociais são constantemente criticadas pela opacidade dos algoritmos de moderação de conteúdo, dos sistemas de recomendação que determinam a ordem e o tipo de conteúdo veiculado aos usuários, bem como da aplicação de seus termos de uso. Por isso, é quase unanimidade entre estudiosos do tema que as plataformas devem ter obrigações de transparência sobre as suas decisões. Afinal, sem transparência não há como os usuários exercerem outros direitos procedimentais, como o devido processo legal e isonomia. O aperfeiçoamento da transparência, contudo, envolve complexos desafios. Fornecer informações detalhadas sobre as decisões de moderação de conteúdo pode acabar tendo efeito reverso, provendo subsídios a pessoas mal intencionadas que terão,

PUC-Rio - Certificação Digital Nº 2112697/CA

com eles, capacidade de burlar a forma como as regras são aplicadas, o que é conhecido como *gaming the system*. Ademais, exigências de transparência podem causar impactos sobre outros interesses relevantes, como a garantia da privacidade. Além disso, existe a complexidade na definição de quais informações devem ser públicas e como elas devem ser apresentadas, contextualizadas e acessadas. Acreditar que quanto mais informação, melhor, não é correto, pois exigências de transparência excessivas incluem custos de oportunidades e diversos riscos.

15. Há, ainda, o desafio da legitimidade das decisões das plataformas. Exemplos de decisões editoriais demonstram que grandes plataformas de redes sociais alcançaram a difícil posição de árbitros na aplicação de suas próprias regras na seara de debates públicos, e no político e eleitoral, em especial, definindo o que pode ou não circular nas redes. Ocorre que o processo decisório das plataformas ainda é opaco, o que acaba dando margem para que as decisões sejam percebidas como arbitrárias ou enviesadas. Isso, somado a um histórico de inconsistência nas decisões e de idas e vindas nas formulações de regras, contribuíram para o desafio da legitimidade das decisões tomadas pelas plataformas.

16. Diante dos desafios apresentados acima, é possível concluir que governos não devem aplicar apenas regras locais para regular o discurso *online,* que se faz presente de forma global, pelas razões a seguir. Em primeiro lugar, a regulação estatal não conseguirá gerenciar a velocidade e a escala em que as plataformas têm que decidir difíceis questões. Em segundo lugar, ainda que fosse possível na prática que regulações estatais gerenciassem os desafios de escala e do volume de conteúdo das redes sociais, a noção de governos terem esse envolvimento na regulação do discurso estaria em tensão com os propósitos democráticos da liberdade de expressão, podendo configurar censura estatal. Em terceiro lugar, há uma tendência de as plataformas removerem em excesso se tiverem risco de serem responsabilizadas pela existência de determinado conteúdo tido por ilícito pelo governo. Em quarto lugar, deveres regulatórios muito exigentes também podem acabar criando ainda mais concentração, por exemplo, ao estabelecerem exigências que apenas as empresas de capital bilionário serão capazes de atender. Por fim, regulações estatais definindo os limites da liberdade de expressão podem colocar em perigo o projeto de plataformas globais, que atuam em escala.

PUC-Rio - Certificação Digital Nº 2112697/CA

17. No contexto em que tanto a autorregulação como a regulação estatal não se mostram suficientes para lidar com os desafios da moderação de conteúdo, a autorregulação regulada se apresenta como uma alternativa promissora à governança da internet. O modelo da autorregulação regulada recorre à cooperação entre Estado e atores regulados para melhor explorar os conhecimentos dos agentes privados, ao mesmo tempo em que garante proteção a direitos fundamentais e valores de interesse público.

18. A regulação estatal da moderação de conteúdo deve se concentrar não na licitude ou ilicitude do conteúdo, mas sim em procedimentos adequados para legitimar as decisões sobre moderação de conteúdo. Esse tipo de regulação não deve pretender vedar a moderação de conteúdo feita com base nos termos de uso das plataformas, pelo contrário. É preciso preservar a autonomia dessas empresas privadas para definirem o tipo de ambiente digital que pretendem fornecer aos usuários, com base na liberdade de iniciativa e de expressão das próprias plataformas, desde que forneçam informações necessárias e suficientes aos usuários, para que entendam o que pode ou não ser feito *online*.

19. Para garantir que as plataformas moderem o discurso de forma sistemática e não arbitrária, a regulação estatal deve se concentrar em impor obrigações procedimentais, como a de divulgação de informações sobre a base da tomada de decisão da plataforma, bem como exigir mecanismos internos de supervisão que tornem as decisões da plataforma publicamente responsabilizáveis. As plataformas devem ser exigidas a mostrar o que estão moderando de acordo com suas regras públicas. As regras devem ser transparentes para trazer maior *accountability* às plataformas.

20. É importante que as decisões das plataformas sobre moderação de conteúdo sejam fiscalizadas por órgãos independentes, de modo a preservar a liberdade de expressão, diante da centralidade das plataformas na esfera pública. Tais órgãos devem servir como fiscalizadores do funcionamento do sistema de deveres procedimentais proposto nesse estudo. Assim, deverão fiscalizar se as plataformas estão cumprindo com os requisitos de transparência, devido processo legal e isonomia, e os deveres mínimos para combater ou minimizar os impactos de conteúdo danoso.

21. O CGI.br tem capacidade de atuar como órgão supervisor do sistema de autorregulação regulada no Brasil. Trata-se de órgão já existente, o que reduziria os

PUC-Rio - Certificação Digital Nº 2112697/CA

custos para adaptação da estrutura, sendo uma organização sem personalidade jurídica responsável por coordenar e atribuir endereços IP em território nacional e por coordenar e registrar nomes de domínio usando o ".br". É, ainda, formado por membros do governo, em número minoritário, e por representantes do setor empresarial, acadêmico, terceiro setor e comunidade científica e tecnológica, tratando-se, portanto, de órgão especializado, independente, e reconhecido internacionalmente.

22. Como a moderação de conteúdo é uma atividade necessária para preservar a liberdade de expressão *online* e não é proibida pelo Marco Civil da Internet, plataformas não podem ser responsabilizadas civilmente pela moderação de conteúdo feita com base nas regras internas, salvo se a medida obedeceu as regras procedimentais a serem impostas por lei, como a devida transparência e direito de resposta do usuário. Apenas em casos de descumprimento das regras procedimentais é que as plataformas poderão ser responsabilizadas pela moderação de conteúdo feita com base em suas regras internas.

PUC-Rio - Certificação Digital Nº 2112697/CA

REFERÊNCIAS BIBLIOGRÁFICAS

ABBOUD, Georges; NERY, Nelson Jr. e CAMPOS, Ricardo. **Fake News e Regulação**. 3ª Edição, ver. e ampl. São Paulo: Thompson Reuters Brasil, 2021.

AFFONSO SOUZA, Carlos. **Bolsonaro edita decreto para acelerar liberação de emendas às vésperas da eleição**. Disponível em: <https://www1.folha.uol.com.br/mercado/2022/09/bolsonaro-edita-decreto-para-acelerar-liberacao-de-emendas-as-vesperas-da-eleicao.shtml>. Acesso em: 12 out. 2022.

_____. **Decreto de Bolsonaro inverte lógica ao impedir moderação de contas e criar indez do que pode ser removido na internet**. Folha de São Paulo, 20 mai. 2021. Disponível em: <https://www1.folha.uol.com.br/poder/2021/05/decreto-de-bolsonaro-inverte-logica-ao-impedir-moderacao-de-contas-e-criar-index-do-que-pode-ser-removido-na-internet.shtml>. Acesso em: 10 out. 2022.

AGÊNCIA BRASIL. Facebook remove 2,5 milhões de posts com discurso de ódio em 6 meses. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2018-05/facebook-remove-25-milhoes-de-posts-com-discurso-de-odio-em-6-meses>. Acesso em: 24 nov. 2022.

AGORA. Assassino nos EUA mata idoso ao vivo no Facebook. Disponível em: <https://agora.folha.uol.com.br/mundo/2017/04/1876437-assassino-nos-eua-mata-idoso-ao-vivo-no-facebook.shtml>. Acesso em: 22 nov. 2022.

AYRES, Ian; BRAITHWAITE, John. **Responsive Regulation: Transcending the Deregulation Debate**. Nova Iorque: Oxford University Press, 1992.

BALKIN, Jack M. **Free Speech is a Triangle**. Columbia Law Review, v. 118, n. 07, p. 2011/2056, 2018.

_____. **Old-school/New-school speech regulation**. Harvard Law Review, Forthcoming, Yale Law School, Public Law Research Paper No. 491, 2014.

BARROSO, Luís Roberto. **Colisão entre liberdade de expressão e Direitos da personalidade. Critérios de ponderação. Interpretação Constitucionalmente adequada do Código Civil e da Lei de Imprensa**. Revista de Direito Administrativo, v. 235, p. 1-36, 2001.

BARROSO, Luna van Brussel. **Liberdade de expressão e democracia na Era Digital: o impacto das mídias sociais no mundo contemporâneo**. Belo Horizonte: Fórum, 2022.

BBC News. Online Safety Bill: New offences and tighter rules. Disponível em: <https://www.bbc.com/news/technology-59638569>. Acesso em: 10 out. 2022.

BINEMBOJM, Gustavo. **Poder de polícia, ordenação, regulação**. 2. ed. Belo Horizonte: Fórum, 2017.

PUC-Rio - Certificação Digital Nº 2112697/CA

BRASIL. Senado Federal. Projeto de Lei N° 2630, de 2020. Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&ts=1648639813988&disposition=inline>. Acesso em: 10 out. 2022.

BRASIL. STF, ADI nº 6.991, Rel. Min. Rosa Weber. j. em 14/09/2021.

BRASIL. STF, Rcl. nº 38.782, Rel. Min. Gilmar Mendes. j. em 23/02/2021.

BRASIL. STF, SL nº 1.248, Rel. Min. Dias Toffoli. j. em 10/09/2019.

BRASIL. STF, ADPF nº 130/DF, Rel. Min. Carlos Ayres Britto, j. em 30/4/2009.

BRASIL. STF, ADPF nº 187/DF, Rel. Min. Celso de Mello, j. em 15/6/2011.

BRASIL. STF, ADPF nº 187/DF, Rel. Min. Celso de Mello, j. em 15/6/2011.

BRASIL. STF, ADPF nº. 187/DF, Rel. Min. Celso de Mello, j. 15/06/2011.

BRASIL. STF, RE nº 685.493, Rel. Min. Marco Aurélio, j. em 20/11/2014.

BRASIL. STJ, AREsp 956.396/MG, Rel. Min Ricardo Villas Bôas Cueva, j. em 17/10/2017.

BRASIL. STJ, RE nº 1.010.606, Rel. Min Dias Toffoli. DJe 19/05/2021.

BRASIL. STJ, RESP 1.193.764/SP, Rel. Min. Nancy Andrighi, j. em 14/12/2010.

BRASIL. STJ, REsp 1.642.997, Rel. Min. Nancy Andrighi, j. em 12/09/2017.

BRASIL. STJ, REsp, 1.186.616/MG, Rel. Min. Nancy Andrighi, j. em 23/08/2011.

BRASIL. STJ, REsp, 1.568.935, Rel. Min. Ricardo Villas Bôas Cueva, j. em 05/04/2016.

BRASIL. TJSP, Apelação Cível 1105667-22.2018.8.26.0100, Rel. Des. Morais Pucci, 35ª Câmara de Direito Privado, j. em 27/07/2020.

BRASIL. TJSP, Apelação Cível 1002982-60.2019.8.26.0565, Rel. Des. Silvério da Silva, 8ª Câmara de Direito Privado, j. em 05/07/2020.

BRASIL. TJSP, Apelação Cível 1004264-50.2014.8.26.0132, Rel. Des. Francisco Loureiro, 1ª Câmara Reservada de Direito Privado, j. em 08/07/2016.

BRASIL. TJSP, Apelação Cível nº 1058263-04.2020.8.26.0100, Rel. Des. Edgar Rosa, 22ª Câmara de Direito Privado, j. em 03/03/2021.

BRASIL. TJSP, Apelação Cível nº 1073111-59.2021.8.26.0100, 6ª Câmara de Direito Privado, Rel. Des. Marcus Vinicius Rios Gonçalves, j. em 26/05/2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

BRASIL. TJMG, Processo nº 5000133-18.2019.8.13.0045, 2º Juizado Especial Cível da Comarca de Caeté, MG, j. em 17/01/2020.

BRASIL. TJSC, Apelação Cível nº 000047-46.2016.8.24.0175, 3ª Câmara de Direito Civil, j. em 06/02/2018.

CALIXTO, Marcelo Junqueira. Desindexação total e parcial nos motores de busca. *In*: SCHREIBER, Anderson, et al. **Direitos fundamentais e sociedade tecnológica**. Indaiatuba, SP: Editora Foco, 2022.

CAMPO, Augustina Del, et. al. **Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet**. Abril, 2021. Disponível em: <https://www.alsur.lat/sites/default/files/2021-06/Responsabilidad%20de%20intermediarios%20PT.pdf>. Acesso em: 10 nov. 2022.

CAMPOS, Ricardo. *et al*. **Regulação de "fake news" no Brasil**. Disponível em <https://institutolgpd.com/wp-content/uploads/2021/10/Regulacao-de-FAKE-NEWS-no-Brasil.pdf. Acesso em 15 nov. 2022.

CASTELLS, Manuel. **Redes de indignação e esperança. Movimentos sociais na era da interne**t. Tradução: Carlos Alberto Medeiros. Rio de Janeiro: Jorge Zahar, 2013.

_____. **Ruptura: A crise da democracia liberal**. Tradução: Joana Angélica D'Avila Melo. Rio de Janeiro: Zahar, 2018.

CELESTE, Edoardo. **Digital Constitutionalism: mapping the constitutional responses to digital technology's challenges**. HIIG Discussion Paper Series No. 2018-02 (2018).

Comissão Europeia. Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Bruxelas, 2020. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825>. Acesso em: 10 out. 2022.

COMITÊ DE SUPERVISÃO. Disponível em: <https://www.oversightboard.com/>. Acesso em: 18 out. 2022.

Consumidor Moderno. Como as redes sociais estão combatendo fake news sobre o coronavirus. Disponível em: <https://www.consumidormoderno.com.br/2020/04/01/redes-sociais-combatendo-desinformacao-coronavirus/>. Acesso em: 12 nov. 2022.

Cornell Law School. 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material. Disponível em: <https://www.law.cornell.edu/uscode/text/47/230>. Acesso em: 8 Jan. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

CUBBY, Inc. v. CompuServe, Inc. U.S. District Court for the Southern District of New York - 776 F. Supp. 135 (S.D.N.Y. 1991) October 29, 1991.

DIETER Grimm. Regulierte Selbstregulierung in der Tradition des Verfassungsstaates, in: Die Verwaltung. Zeitschrift fur Verwaltungrecht und Verwaltungswissenschaften, Caderno 4, Regulierte Selbstregulierung als Steuerungskonzept des Gewahrleistungsstaates, Duncker & Humblot, Berlim, 2001, p. 18. *In*: ABBOUD, Georges, NERY junior, Nelson; CAMPOS, Ricardo. **Fake News e regulação**, São Paulo: Thomson Reuters Brasil, 2021.

DOUEK, Evelyn. **Facebook's Oversight Board: move fast with stable infrastructure and humility**. North Carolina Journal of Law and Technology Volume 21 (2019).

_____. **Governing online speech**. Columbia Law Review, v. 121, n. 03, 2021.

_____. **Verified Accountability: Self-regulation of content moderation as an answer to the special problems of speech regulation**. Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903.

EDWARDS, Lillian; VEALE, Michael. "**Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For**." Duke Law & Technology Review 16, no. 1 (2017): 67.

ESTARQUE, Marina; ARCHEGAS, João Victor; BOTTINO, Celina; PERRONE, Christian. **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/>. Acesso em: 07 jul. 2021.

EXAME. Facebook exclui usuário que postou 'A Origem do Mundo', de Courbet. Disponível em: <https://exame.com/tecnologia/facebook-exclui-usuario-que-postou-a-origem-do-mundo-de-courbet/>. Acesso em: 20 nov. 2022.

FEATURING MONIKA BICKERT & JONATHAN ZITTRAIN IN CONVERSATION. Disponível em: <https://cyber.harvard.edu/events/state-online-speech-and-governance>. Acesso em: 12 nov. 2022.

FONSECA, Gabriel; e VERONESE, Alexandre. Desinformação, fake news e mercado único digital: a potencial convergência das políticas públicas da União Europeia com os Estados Unidos para melhoria dos conteúdos comunicacionais. *In*: **Fake News e as eleições 2018**. Cadernos Adenauer, 2018.

FORTUNE. Check Out Alphabet's New Tool to Weed Out the 'Toxic' Abuse of Online Comments. Disponível em: <https://fortune.com/2017/02/23/alphabet-jigsaw-perspective-comment-moderator/>. Acesso em: 12 nov. 2022.

FRAZÃO, Ana. **Plataformas Digitais e o Negócio de Dados: Necessário Diálogo entre o Direito da Concorrência e a Regulação dos Dados**. RDP, Brasília, Volume 17, n. 93, 58-81, maio/jun. 2020.

PUC-Rio - Certificação Digital Nº 2112697/CA

FRAZÃO, Ana; MEDEIROS, Ana Rafaela. Responsabilidade civil dos provedores de internet: A liberdade de expressão e o art. 19 do Marco Civil. *In*: JÚNIOR; Marcos Ehrhardt; LOBO, Fabíola Albuquerque; e ANDRADE, Gustavo. **Liberdade de expressão e relações privadas**. Editora Fórum, e-book.

FUKUYAMA, Francis, et al. **Report of the working group on platform scale**. Stanford Program on Democracy and the Internet, 2020. Disponível em: <https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale>. Acesso em: 12 nov. 2022.

Germany fines Facebook for under-reporting complaints. Reuters, 2 jul. 2019. Disponível em: <https://www.reuters.com/article/us-facebook-germany-fine-idUSKCN1TX1IC>. Acesso em: 22 nov. 2022.

GILLESPIE, Tarleton. **Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media**. New Haven, CT: Yale University Press, 2018.

Google in Europe. Partnering to help curb the spread of terrorist content online . Disponível em: <https://blog.google/around-the-globe/google-europe/partnering-help-curb-spread-terrorist-content-online/>. Acesso em: 12 nov. 2022.

GOV.UK. Disponível em: <https://www.gov.uk/government/consultations/online-harms-white-paper>. Acesso em: 10 out. 2022.

HUMAN RIGHTS COMMITTEE. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. 11 mai. 2016. UN Doc A/HRC/32/38. Disponível em: <https://undocs.org.en/A/HRC/32/38>. Acesso em: 12 nov. 2022.

KADRI, Thomas e KLONICK, Kate. **Facebook v. Sullivan: public figures and newsworthiness in online speech**. Southern California Lay Review Volume 93 (2019).

KAYE, David**. Speech Police: the global struggle to govern the internet**. Columbia Global Reports, 2019.

KELLER, Clara Iglesias. **Regulação nacional de serviços na Internet: exceção, legitimidade e o papel do Estado**. Tese (Doutorado), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019.

KLONICK, Kate. "**The New Governors: The People, Rules, and Processes Governing Online Speech**," Harvard Law Review 131 (2018): 1598.

KOSSEF, Jeff. **The Twenty-Six Words That Created the Internet**. Cornell University Press, 2019, capítulo 12.

_____. **A user's guide to Section 230, and a legislator's guide to amending it (or not)**. Berkeley Technology Law Journal, v. 37, nº 2, 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

KRAUS, Mariella; PANSIERI, Flávio; PAVAN, Stefano Ávila. **Desinformação, pós-verdade e democracia**: **Uma análise no contexto do Estado democrático de direito**. Revista Jurídica Unicuritiba. Curitiba.V.04, n.66, p.163-196 [Received/Recebido: Maio 23, 2021; Accepted/Aceito: Julho 23, 2021].

KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim; VIEIRA, Victor Barbieri Rodrigues. **Transparência na moderação de conteúdo: tendências regulatórias nacionais**. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: <https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf>. Acesso em: 19 out. 2022.

LEERSEEN, Paddy. **The soap box as a black box: regulating transparency in social media recommender systems**. European Journal of Law and Technology, v. 11, n. 2, 2020.

LESSIG, Lawrence. **Code: version 2.0**, Basic Books, 2006.

LONGHI, João Victor Rozatti. **Marco Civil da Internet no Brasil: Breves considerações sobre seus fundamentos, princípios e análise crítica do regime de responsabilidade dos provedores**. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4635703/mod_resource/content/1/capi%CC%81tulo%205%20DIREITO%20PRIVADO%20E%20INTERNET.pdf>. Acesso em: 23 nov. 2022.

MANILA PRINCIPLES on intermediary liability. Disponível em: <https://manilaprinciples.org/pt-br.html>. Acesso em: 18 out. 2022.

MARSHALL, William P. The Truth Justification for Freedom of Speech. *In*: STONE, Adrienne; SCHAUER, Frederick (EDs.). **Freedom of Speech**. United Kingdom: Oxford University Press, 2021.

MARTINS, Guilherme Magalhães . **Vulnerabilidade e responsabilidade civil na Internet: a inconstitucionalidade do Artigo 19 do Marco Civil**. Revista de Direito do Consumidor , v. 137, p. 33-59, 2021.

_____. **Responsabilidade objetiva do provedor de aplicações de internet**. Disponível em: <https://www.conjur.com.br/2015-nov-18/guilherme-martins-responsabilidade-objetiva-provedor-internet>. Acesso em: 23 nov. 2022.

MENDONÇA, Eduardo. **Retrocesso autoritário**. Estadão, jun. 2021. Disponível em: <https://politica.estadao.com.br/blogs/fausto-macedo/retrocesso-autoritario/>. Acesso em: 10 out. 2022.

META. Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process. Disponível em: <https://about.fb.com/news/2018/04/comprehensive-community-standards/>. Acesso em: 20 nov. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

MIRAGEM, Bruno. **Responsabilidade por danos na sociedade de informação e proteção do consumidor: desafios atuais na regulação jurídica da internet**. Revista de Direito do Consumidor, São Paulo, v. 70, p. 1-42, abr. 2009.

MONIKA, Bickert. "Defining the boundaries of free speech on social media". *In*: **The free speech century**. Edited by Lee C. Bollinger and Geoffrey R. Stone. 2019.

MORAES, Maria Celina Bodin de; TEFFÉ, Chiara Spadaccini de. **Redes sociais virtuais: privacidade e responsabilidade civil**. **Análise a partir do Marco Civil da Internet**. Pensar, Fortaleza, v. 22, n. 1, p. 108-146, jan./abr. 2017.

MULHOLLAND, Caitlin. Responsabilidade civil indireta dos provedores de serviço de Internet e sua regulação no Marco Civil da Internet. *In*: CELLA, José Renato Gaziero; NASCIMENTO, Aires Jose Rover, Valéria Ribas do. (orgs). **Direito e novas tecnologias**. 1ª Ed. Florianópolis: CONPEDI, 2015, v. 1.

NITRINI, Rodrigo Vidal. **Liberdade de expressão nas redes sociais**. **O problema jurídico da remoção de conteúdo pelas plataformas**. Belo Horizonte: Editora Dialética, 2021.

OFICINADANET. Disponível em: <https://www.oficinadanet.com.br/post/16064-quais-sao-as-dez-maiores-redes-sociais>. Acesso em: 20 nov. 2022.

OPENDEMOCRACY. Disponível em: <https://www.opendemocracy.net/pt/censura-twitter-facebook-donald-trump-consequencia-democracia/>. Acesso em: 12 nov. 2022.

OSORIO, Aline. **Direito eleitoral e liberdade de expressão**. Belo Horizonte: Fórum, 2017.

Portuguese The Santa Clara Principles. Disponível em: <https://santaclaraprinciples.org/pt/cfp/>. Acesso em: 18 out. 2022.

Protecting Election Integrity. Facebook, [s.d]. Disponível em: <https://www.facebook.com/business/m/election-integrity>. Acesso em: 20 nov. 2022.

QUEIROZ, João Quinelato de. **Responsabilidade Civil na Rede: danos e liberdade à luz do marco civil da internet**. 1. ed. Rio de Janeiro: Processo, 2019. v. 1.

RAMOS, Carlos Eduardo Vieira. **O direito das plataformas Digitais - Regulação Privada da Liberdade de Expressão na Internet - Procedimento, Legitimidade e Constitucionalização**. Curitiba: Juruá, 2021, p. 24.

RECUERO, Raquel. Rede social. *In*: AVORIO, A.; SPYER, J. (Org.). **Para entender a Internet**. Versão rev. e ampl., 2015.

RHEINGOLD, Howard. **The Virtual Community**. Disponível em: <https://www.rheingold.com/vc/book/10.html>. Acesso em: 12 nov. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

ROSEN, Jeffrey. "**The Deleted Squad**", artigo publicado pelo The New York Republic, em 29/04/2013.

SARMENTO, Daniel. A Liberdade de Expressão e o Problema do "Hate Speech". *In:* **Livres e iguais: estudos de Direito Constitucional**. Rio de Janeiro: Lumen Juris, 2006.

SCHREIBER, Anderson. **Marco Civil da Internet: avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro**. Disponível em: <andersonschreiber.com.br/downloads/artigo-marco-civil-internet.pdf>. Acesso em: 8 Jan. 2022.

SOUZA, Carlos Affonso Souza; TEFFÉ, Chiara Spadaccini de. **Responsabilidade dos provedores por conteúdos de terceiros na internet**. Disponível em: <https://www.conjur.com.br/2017-jan-23/responsabilidade-provedor-conteudo-terceiro-internet>. Acesso em: 8 Jan. 2022.

STRATTON OAKMONT, Inc. v. Prodigy Services Co. Supreme Court, Nassau County, New York, Trial IAS Part 34. May 24, 1995.

STROPPA, Tatiana, et al. **A seção 230 do CDA e o artigo 19 do Marco Civil da Internet.** Disponível em: <https://www.conjur.com.br/2022-mai-04/direito-digital-secao-230-cda-artigo-19-marco-civil-internet#_ftn6>. Acesso em: 23 nov. 2022.

SUNDFELD, Carlos Ari; ROSILHO, André. **A governança não estatal da internet e o direito brasileiro**. RDA, Rio de Janeiro, v. 270, p. 41/79, set./dez. 2015.

SUNSTEIN, Cass. **#Republic: divided democracy in the age of social media**. Princeton: Princeton University Press, 2017.

SUZOR, Nicolas. **Lawless: the secret rules that govern or digital lives**, Cambridge University Press, 2019.

TEFFÉ, Chiara Spadaccini de. Exposição não consentida de imagens íntimas: como o direito pode proteger as mulheres? *In*: ROSENVALD, Nelson; DRESCH, Rafael de Freitas Valle; WESENDONCK, Tula (coord.). **Responsabilidade civil: novos riscos**. Indaiatuba: Foco, 2019.

The Digital Millennium Copyright Act Of 1998. Disponível em: <https://www.copyright.gov/legislation/dmca.pdf>. Acesso em: 8 Jan. 2022.

The Guardian. Disponível em: <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>. Acesso em: 20 nov. 2022.

PUC-Rio - Certificação Digital Nº 2112697/CA

THE NEW YORK TIMES. Disponível em: <https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html>. Acesso em: 23 nov. 2022.

The Risk of Racial Bias in Hate Speech Detection. Disponível em: <https://aclanthology.org/P19-1163.pdf>. Acesso em: 12 nov. 2022.

The Washington Post. Disponível em: <https://www.washingtonpost.com/news/the-switch/wp/2016/07/07/why-facebook-took-down-the-philando-castile-shooting-video-then-put-it-back-up/>. Acesso em: 22 nov. 2022.

TSE. Contra fake news, Instagram e Facebook colocam avisos em postagens sobre Eleições 2022. Disponível em: <https://www.tse.jus.br/comunicacao/noticias/2021/Dezembro/contra-fake-news-instagram-e-facebook-colocam-avisos-em-postagens-sobre-eleicoes-2022>. Acesso em: 12 nov. 2022.

TWOREK, heidi, LEERSEN, Paddy. **An analysis of Germany's NetzDG Law**. Artigo publicado pelo Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression - Institute for Information Law (Universiteit Van Amesterdam), em abril de 2019.

VALENTE, Mariana G, MONTEIRO, Artur Pericles Lima, CRUZ, Francisco Brito, DA SILVEIRA, Juliana Fonteles. **Armadilhas e caminhos na regulação da moderação de conteúdo, Diagnósticos & Recomendações** (São Paulo: InternetLab, 2021). Disponível em: <https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf>. Acesso em: 10 nov. 2022.

VENTURI, Thaís G. Pascoaloto. **Redes Sociais: Platforms ou Publishers? - Parte I**. Disponível em: <https://www.migalhas.com.br/coluna/direito-privado-no-common-law/339965/redes-sociais-platforms-ou-publishers--parte-i>. Acesso em: 23 nov. 2022.

WIKINOTÍCIAS. França: Conselho Constitucional censura a lei Avia. Disponível em: <https://pt.wikinews.org/wiki/Fran%C3%A7a:_Conselho_Constitucional_censura_a_lei_Avia>. Acesso em: 10 out. 2022.

WIRED. Disponível em: <https://www.wired.com/story/facebook-content-standards-politicians-exemption-dave-willner/>. Acesso em: 12 nov. 2022.

WU, Tim. Is the First Amendment Obsolete? *In*: POZEN, David E. (Ed.). **The Perilous Public Square**, New York: Columbia University Press, E-book Kindle. (Não paginado).

_____. **The Attention Merchants**: **The Epic Scramble to Get Inside Our Heads**. 21. New York, NY. 2019.

PUC-Rio - Certificação Digital Nº 2112697/CA

Youtube Help. How Content ID works. Disponível em: <https://support.google.com/youtube/answer/2797370?hl=en>. Acesso em: 12 nov. 2022.

ZUBOFF, Shoshana. **A era do capitalismo de vigilância: a luta por um futuro humano na nova fronteira do poder**; tradução George Schlesinger. 1ª ed. Rio de Janeiro: Intrínseca, 2020.

ZUCKERBERG, Mark. "**The internet needs new ruled. Let's start in these four areas**", artigo publicado pelo The Washington Post, em 30/03/2019.

ZURTH, Patrick. **The German NetzDG as role model or cautionary tale?** Implications for the debate on social media Liability. 31 Fordham Intell, Prop, Media & Ent. L.j. 1084, p. 1130, 2021.

PUC-Rio - Certificação Digital Nº 2112697/CA