Marcelo Brígido Ayala Pereira

MODERAÇÃO DE CONTEÚDO NAS REDES SOCIAIS: uma proposta de regulação por normas procedimentais

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

Prof.ª Dr.ª Caitlin Sampaio Mulholland
Orientadora
Departamento de Direito – PUC-Rio

Rio de Janeiro
Abril de 2025

Marcelo Brígido Ayala Pereira

MODERAÇÃO DE CONTEÚDO NAS REDES SOCIAIS: uma proposta de regulação por normas procedimentais

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Mestrado em Direito Civil Contemporâneo e Prática Jurídica da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Aprovada pela Comissão Examinadora abaixo:

Prof.ª Dr.ª Caitlin Sampaio Mulholland
Orientadora
Departamento de Direito – PUC-Rio

Prof. Dr. Carlos Nelson Konder
Departamento de Direito – PUC-Rio

Prof. Dr. Carlos Affonso de Souza
Departamento de Direito – PUC-Rio

Rio de Janeiro, 07 de abril de 2025

Marcelo Brígido Ayala Pereira
Mestre em Direito Civil pela PUC-Rio

Ficha Catalográfica

Pereira, Marcelo Brígido Ayala

Moderação de conteúdo nas redes sociais : uma proposta de regulação por normas procedimentais / Marcelo Brígido Ayala Pereira; orientadora: Caitlin Sampaio Mulholland. – 2025.
188 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Direito, 2025.
Inclui bibliografia

1. Direito – Teses. 2. Liberdade de expressão. 3. Moderação de conteúdo. 4. Redes sociais. 5 Regulação. 6. Autorregulação. 7. Autorregulação regulada. I. Mulholland, Caitlin. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Direito. III. Título.

CDD: 340

Agradecimentos

O agradecimento é o mínimo que se espera de quem alcança um objetivo, ciente de que, ao longo do caminho, contou com pessoas dispostas a prestar apoio de diversas maneiras. Ao longo deste trabalho de pesquisa, as contribuições recebidas foram peças chaves para a conclusão desta dissertação.

Ao meu querido pai, Marcio, por ter sido o principal incentivador nesta jornada, sempre me motivando a buscar novos desafios. Você foi a minha fortaleza do início ao fim deste projeto. Mesmo na ausência de sua presença física durante os momentos finais de elaboração deste trabalho, minha maior motivação sempre foi transformá-lo em uma conquista nossa. Agora, ao chegar a este momento, firme na convicção de que você está ao meu lado, dedico-lhe este trabalho.

À minha mãe, Elizabeth, pelo amor incondicional, pela proteção e pelo suporte em todas as etapas da minha vida pessoal, profissional e acadêmica. Agradeço por depositar tanta energia e dedicação no meu futuro.

À minha namorada, Roberta, pela parceria, carinho e companheirismo durante este percurso, repleto de finais de semana, feriados e noites de estudo. Não poderia deixar de mencionar suas sugestões, sua paciência e a revisão do texto final da dissertação.

Ao amigo Patrick, pelas inúmeras contribuições feitas ao longo da pesquisa. Nossas discussões sobre temas controvertidos, metodologias de pesquisa e trocas de ideias e reflexões, em especial diante do dinamismo da matéria – à medida que novas notícias sobre os atuais conflitos oriundos da ausência de regulação das redes sociais surgiam –, foram fundamentais para o bom desenvolvimento do trabalho.

À professora Caitlin Mulholland, pela dedicada orientação ao longo de todo o trabalho de pesquisa, bem como pelo excepcional empenho no aconselhamento, na disponibilização de referências bibliográficas e na correção da dissertação.

Aos professores Carlos Konder e Carlos Affonso, pelas valiosas contribuições durante a qualificação do projeto de pesquisa e pela indicação de referências bibliográficas.

Resumo

PEREIRA, Marcelo Brígido Ayala. **Moderação de conteúdo nas redes sociais: uma proposta de regulação por normas procedimentais**. Orientadora: Caitlin Sampaio Mulholland. Rio de Janeiro, 2025, 188 p. Dissertação de Mestrado – Departamento de Direito, Pontifica Universidade Católica do Rio de Janeiro

Esta dissertação analisa a moderação de conteúdo nas redes sociais e a necessidade de sua regulação no Brasil. Inicialmente, examina-se a evolução das tecnologias de comunicação e como as redes sociais criaram uma nova esfera pública. A pesquisa aborda a disseminação de *fake news* e a demanda por regulação, analisando a evolução histórica da liberdade de expressão e a falta de regras específicas sobre moderação no Marco Civil da Internet – MCI. O estudo também destaca a atuação do Supremo Tribunal Federal – STF na análise da (in)constitucionalidade do artigo 19 do MCI. Os desafios de transparência e *accountability* nas redes sociais são discutidos, assim como a capacidade dessas plataformas de atendê-los. Em seguida, são comparados os modelos de regulação: autorregulação, heterorregulação e autorregulação regulada. Propõe-se um modelo de autorregulação regulada, que busca combinar a intervenção estatal com a expertise das plataformas, para harmonizar a liberdade de expressão, a liberdade editorial e a proteção de interesses públicos no ambiente *online*, através de normas procedimentais. Ao final, recomendações de organizações nacionais e internacionais são analisadas para estabelecer as melhores regras e práticas globais, visando a implementação de um procedimento adequado para a moderação de conteúdo, cuja supervisão poderá ser feita por órgão fiscalizador independente.

Palavras-chave

Liberdade de expressão. Moderação de Conteúdo. Redes sociais. Regulação. Autorregulação. Autorregulação regulada.

Abstract

PEREIRA, Marcelo Brígido Ayala. **Social media content moderation: a proposal for procedural regulation**. Advisor: Caitlin Sampaio Mulholland. Rio de Janeiro, 2025, 188 p. Master ’s Thesis – Department of Law, Pontifical Catholic University of Rio de Janeiro

This dissertation analyzes content moderation on social media and the need for its regulation in Brazil. Initially, it examines the evolution of communication technologies and how social media has created a new public sphere. The research addresses the dissemination of fake news and the demand for regulation, analyzing the historical evolution of freedom of expression and the lack of specific rules regarding content moderation in the Brazilian Civil Rights Framework for the Internet (Marco Civil da Internet – MCI). The study also highlights the role of the Supreme Federal Court – STF in the analysis of the (un)constitutionality of Article 19 of the MCI. The challenges of transparency and accountability on social media will be discussed, as well as the ability of these platforms to fulfill them. Subsequently, regulatory models will be compared: self-regulation, public regulation and corregulation. A model of corregulation is proposed, which aims to combine the state intervention with the platforms' expertise to harmonize freedom of expression, editorial freedom, and the protection of public interests in the online environment through procedural rules. Finally, recommendations from national and international organizations will be analyzed to establish the best global rules and practices aimed at implementing an appropriate procedure for content moderation, which could be overseen by an independent regulatory body.

Keywords

Freedom of Speech. Content Moderation. Social Media. Regulation. Self-regulation. Corregulation.

Sumário

INTRODUÇÃO: DE QUEM É O PASSADO, O PRESENTE E O FUTURO?

"Quem controla o passado, controla o futuro; quem controla o presente, controla o passado" é o *slogan* do Partido, uma instituição que manipula a verdade, no romance *1984*, de George Orwell.[1] No cenário distópico do autor, o Ministério da Verdade – um órgão administrado pelo Partido – reescreve a história para controlar o debate sobre o presente e o futuro; limita as discussões no presente para moldar a percepção do futuro e apagar o passado; e constrói um futuro que justifique o passado reconfigurado.

A reflexão de Orwell aborda os riscos do controle da liberdade de expressão pelo Estado, historicamente visto como um potencial censor da livre circulação de informações e ideias, por motivações políticas e morais. A lição ensinada é que o debate sobre a moderação estatal da liberdade de expressão deve ser tratado com cautela, a fim de evitar ou, de maneira pragmática, minimizar ao máximo violações aos direitos humanos — incluindo a liberdade de expressão.

O futuro distópico apresentado por Orwell, no entanto, não previu a crescente influência e protagonismo das empresas privadas com a revolução tecnológica na comunicação. Tampouco a possibilidade de interagir, por meio das redes sociais,[2] com uma comunidade com mais de dois bilhões de pessoas com poucos cliques. A coleta massiva de dados e a capacidade de manipular, por meio de algoritmos, o que uma pessoa consome, são poderes inimagináveis, que o autor não foi capaz de antecipar.

O efeito prático disso é que, no imaginário coletivo, o debate sobre moderação de conteúdo no ambiente *online* foca nos riscos representados pela atuação de alguns atores (principalmente, os Estados), mas ignora o surgimento de outros relevantes agentes, como as empresas de tecnologia, que, com o mesmo

---

[1] ORWELL, George. *1984*. Traduzido por Karla Lima. São Paulo: Principis, 2021, p. 43, *ebook*.

[2] As redes sociais podem ser conceituadas como "plataformas interativas que permitem que usuários montem um *perfil pessoal* e, a partir dele em seu nome, *gerem conteúdos* (tais como texto, postagens, imagens ou vídeos) que não tornam-se visíveis a terceiros, mas que *sirvam de elo para a formação de conexões interpessoais em rede"*. Essa definição, que exclui os aplicativos de mensageria instantânea, como o *Whatsapp*, pode ser complementada ainda pela noção de que as redes sociais "customizam e personalizam a ordenação e a visibilidade de conteúdos aos usuários por meio de algoritmos, de modo que cada perfil tem uma experiência própria de visualização durante seu uso". (NITRINI, Rodrigo Vidal. *Liberdade de Expressão nas Redes Sociais: o problema jurídico da remoção de conteúdo pelas plataformas*. Belo Horizonte: Dialética, 2021, p. 18).

poder transformador, alteraram não apenas a dinâmica do debate público, mas a própria forma como as pessoas interagem, agora quase sempre mediadas pelas redes sociais.[3]

É preciso refletir sobre essa nova realidade. A história brasileira contemporânea demonstrou que a ausência de regulação sobre a disseminação de notícias falsas tem agravado crises institucionais. A propagação massiva de informações falsas durante o Covid-19[4] e a ausência de moderação de conteúdo em torno dos atos de vandalismo ocorridos no dia 8 de janeiro de 2024 em Brasília, foram eventos que impulsionaram o debate legislativo em torno da regulação de redes sociais no Brasil. Esses eventos contribuíram para o início do julgamento, no STF, da constitucionalidade do artigo 19 do Marco Civil da Internet, que garante imunidade à responsabilização das redes sociais por danos decorrentes de conteúdo de terceiros, ressalvada a hipótese de descumprimento de ordem judicial prévia.[5]

Diante desse cenário, emergem alguns dos questionamentos que orientam a pesquisa desta dissertação: a regulação da liberdade de expressão nas redes sociais é necessária no contexto brasileiro? Quais seriam os limites para que a regulação das redes sociais voltada à atividade de moderação de conteúdo das plataformas não se torne um instrumento de censura? Qual seria o modelo mais adequado de regulação das redes sociais capaz de promover a liberdade de expressão e demais direitos dos usuários e, ao mesmo tempo, preservar os direitos das plataformas?

Para responder a esses questionamentos, o estudo parte da premissa de que a liberdade de expressão e a moderação de conteúdo não são conceitos antagônicos, mas devem coexistir dentro de um arcabouço normativo que garanta transparência, previsibilidade e segurança jurídica.

A busca por um modelo regulatório eficaz deve considerar tanto a proteção de direitos individuais dos usuários, prevenindo abusos e manipulações

---

[3] Embora existam distinções técnicas entre plataformas digitais e redes sociais, os termos serão utilizados, em determinados contextos, como sinônimos ao longo deste trabalho de pesquisa, por se tratar de conceitos relacionados, cuja distinção não é capaz de alterar significativamente o tratamento jurídico que será objeto do estudo.

[4] AJZENMAN, Nicolás; CAVALCANTI, Tiago; DA MATA, Daniel. *More than words: leaders' speech and risky behavior during a pandemic*. Cambridge-INET Working Paper Series No. 2020/19; Cambridge Working Papers in Economics: 2034, 2020. Disponível em: <https://econpapers.repec.org/paper/camcamdae/2034.htm>. Acesso em 28 fev. 2025.

[5] GUIDO, Gabriela. *8 de janeiro e atentado a bomba legitimam STF para julgar regulamentação das redes sociais*. Valor, publicado em 27 nov. 2024. Disponível em: <https://valor.globo.com/politica/noticia/2024/11/27/8-de-janeiro-e-atentado-a-bomba-legitimam-stf-para-julgar-regulamentacao-das-redes-sociais.ghtml>. Acesso em 28 fev. 2025.

informacionais nas redes sociais, quanto assegurar a liberdade e a autonomia das plataformas de criarem o ambiente virtual que considerem atrativos aos seus usuários.

A dissertação está estruturada, portanto, em três grandes eixos temáticos. O primeiro capítulo apresenta algumas premissas sobre a liberdade de expressão. Nesse capítulo, será contextualizada a nova esfera pública digital, criada pelas redes sociais, como o novo campo de debate social, onde ideias, discursos político-econômicos e mobilizações sociais se desenrolam de forma global, mas são regidos basicamente por normas privadas.

Em seguida, o trabalho perpassará pela evolução legislativa da liberdade de expressão no Brasil, os impactos das novas tecnologias e os desafios do fenômeno das *fake news*, especialmente considerando o vácuo normativo brasileiro e a atuação do Poder Público na tentativa de "regular" as redes sociais. Além disso, serão abordadas as razões para a regulação, no contexto brasileiro, da moderação de conteúdo nas redes sociais, apresentando, em linhas gerais, os modelos de regulação existentes (autorregulação, heterorregulação ou corregulação).

No segundo capítulo, o foco recairá sobre a regulação privada das redes sociais (autorregulação), sua legitimidade, seus limites e desafios. Será examinado ainda como as próprias plataformas exercem a moderação de conteúdo, seja por meio de controle automatizado (inteligência artificial e algoritmos), bloqueio geográfico, sistemas de sinalização da comunidade (*flagging*) ou revisão humana.

Ainda nesse capítulo, será abordado o conceito de eficácia horizontal dos direitos fundamentais, avaliando como direitos individuais podem ser garantidos mesmo em interações privadas dentro do ecossistema digital, estabelecendo os limites da moderação de conteúdo. Por fim, o capítulo analisará os desafios de transparência e *accountability* das plataformas digitais, destacando a necessidade de mecanismos de fiscalização claros e eficientes que ainda estão por se desenhar no debate nacional.

O terceiro e último capítulo proporá um modelo de regulação bifásica da liberdade de expressão, no qual o Estado não detém o monopólio regulatório, mas atua em parceria com a sociedade civil e as próprias redes sociais. Nesse modelo *policêntrico*, caberá ao Estado estabelecer normas procedimentais que garantam uma moderação de conteúdo proporcional e razoável, enquanto a aplicação material da moderação permanecerão sob responsabilidade das plataformas – sempre sujeita

à revisão posterior pelo Poder Judiciário. Essa colaboração complexa entre o setor privado e o Poder Público configura um regime de autorregulação regulada, no qual devem ser implementadas medidas de transparência, devido processo e isonomia para assegurar um ambiente digital saudável, funcional e eficaz.

Também serão analisadas contribuições de organizações nacionais e internacionais que desenvolveram estudos e propostas sobre normas procedimentais e boas práticas a serem adotadas na moderação de conteúdo, servindo como referência para a regulação da atividade de moderação de conteúdo realizada pelas plataformas digitais no ordenamento jurídico brasileiro.

O produto final desta análise consistirá no estabelecimento de um padrão estruturado de deveres procedimentais a ser observado pelo legislador brasileiro, baseado nas propostas analisadas das referidas organizações e nos ensinamentos doutrinários sobre tema. Acredita-se que esse padrão normativo poderia assegurar a proteção dos direitos humanos e o legítimo exercício da liberdade de expressão dos usuários nas redes sociais, assim como os direitos editorais das plataformas digitais.

Por fim, será analisada a proposta de criação de um órgão fiscalizador independente, responsável por supervisionar a aplicação dessas normas procedimentais pelas plataformas digitais, mitigar falhas sistêmicas e garantir um equilíbrio adequado entre regulação e liberdade de expressão.

1. A LIBERDADE DE EXPRESSÃO NAS REDES SOCIAIS

1.1. O avanço das tecnologias de comunicação e as novas formas de interação social nas redes sociais: uma esfera pública participativa

A sociedade contemporânea está marcada pela globalização e pelos avanços tecnológicos, assim como pelo volume, velocidade, variedade e valor dos dados e informações disseminados nas redes sociais. A expressiva coleta de dados reforça a crescente tendência de tomada de decisões sociais e econômicas por algoritmos, robôs e inteligência artificial. Esses fenômenos são conhecidos como Sociedade de Dados Massivos (*Big Data Society*)[6] ou Algorítmica (*Algorithmic Society*).[7]

O expressivo progresso das tecnologias de comunicação na virada do século XX e para o século XXI impulsionou a criação e a expansão das redes sociais,[8] que propiciam novas formas de interação social. Essas novas formas de interação social impactam cotidianamente a vida das pessoas, com reflexos visíveis sobre aspectos sociais, econômicos e políticos.[9]

As redes sociais descentralizaram o debate público – que até então estava concentrado nas mãos de poucos atores sociais, em especial, dos veículos de comunicação da mídia tradicional –,[10] reduzindo a hierarquia sobre a qual foi edificada a verticalidade do poder público.[11]

Ao permitir que qualquer pessoa possa se manifestar publicamente, por meio de serviços gratuitos fornecidos por plataformas digitais,[12] com capacidade de

---

[6] PEREIRA DE LIMA, Cíntia Rosa; FRANCO DE MORAES, Emanuele Pezati; PEROLI, Kelvin. O necessário diálogo entre o Marco Civil da Internet e a Lei Geral de Proteção de Dados para a coerência do sistema de responsabilidade civil diante das Novas Tecnologias. In: *Responsabilidade civil e novas tecnologias*; coordenado por Guilherme Magalhães Martins. Nelson Rosenvald – Indaiatuba, SP. Editora Foco, 2020, p. 146.

[7] BALKIN, Jack M., *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation* (September 9, 2017). UC Davis Law Review, (2018 Forthcoming), Yale Law School, Public Law Research Paper No. 615, p. 3. Disponível em: <https://ssrn.com/abstract=3038939>. Acesso em 24 nov. 2024.

[8] FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*. São Paulo: Perspectiva, 2020, p. 11.

[9] BARROSO, Luis Roberto. BARROSO, Luna Van Brussel. Prefácio à 3ª Edição. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, p. 6.

[10] BARROSO, Luis Roberto. Prefácio. In: Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*. Belo Horizonte: Fórum, 2022, p. 15.

[11] FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*..., p. 11.

[12] A gratuidade dos serviços das redes sociais tem como revés sérios problemas para a privacidade de dados pessoais de seus usuários. O acesso gratuito ao serviço das plataformas está inserido no

alcance a um número expressivo de pessoas, pluralizando as fontes e multiplicando exponencialmente o volume e o acesso às informações, as redes sociais contemporâneas propiciam, em tese, uma esfera pública participativa.[13]-[14]

A esfera pública é um espaço fundamental para a democracia, onde as pessoas expressam opiniões e compartilham perspectivas sobre questões sociais, que não se limitam a política governamental, abrangendo temas como esportes, cultura, comércio etc. Ela é moldada e governada por instituições – a maioria delas privadas – que a tornam funcional ou disfuncional. Em uma esfera pública digital, as plataformas digitais são as principais instituições capazes de manter ou prejudicar a saúde da esfera pública.[15]

A esfera pública digital criada pelas redes sociais no século XXI se difere daquela que lhe precedeu no século XX, porque quase todo o conteúdo publicado ou transmitido pelas mídias impressa e de radiodifusão (ou audiovisual) era produzido pelas empresas que detinham os meios de comunicação. Essas mídias não permitiam a participação das pessoas na elaboração e divulgação de conteúdo, que desempenhavam o papel de meros expectadores. As redes sociais, por outro lado, hospedam conteúdo criado por muitos usuários, que são tanto expectadores quanto criadores.[16]

---

contexto de um "ecossistema de publicidade digital", que coleta esses dados para exploração comercial, a fim de submeter os usuários à publicidade com alto grau de refinamento (NITRINI, Rodrigo Vidal. *Liberdade de Expressão nas Redes Sociais: o problema jurídico da remoção de conteúdo pelas plataformas...*, p. 15).

[13] BARROSO, Luis Roberto. Prefácio. In: Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 15.

[14] Apesar do papel de destaque das plataformas digitais privadas para a liberdade de expressão e a participação na cultura democrática, Kate Klonick ressalva que a possibilidade de a pessoa publicar livre e instantaneamente nas redes sociais precisa considerar, por outro lado, a atuação proativa das grandes plataformas digitais na curadoria do discurso público, por meio de diversos mecanismos de moderação de conteúdo, cuja operacionalização ainda é insuficientemente conhecida. (KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech* (March 20, 2017). 131 Harv. L. Rev., pp. 1598-1599. Disponível em: <https://ssrn.com/abstract=2937985>. Acesso em 24 nov. 2024).

[15] BALKIN, Jack M. *How to Regulate (and Not Regulate) Social Media* (November 8, 2019). 1 Journal of Free Speech Law 71 (2021), Knight Institute Occasional Paper Series, No. 1 (March 25, 2020), Yale Law School, Public Law Research Paper Forthcoming, pp. 72-73. Disponível em: <https://ssrn.com/abstract=3484114>. Acesso em 29 nov. 2024.

[16] BALKIN, Jack M., *How to Regulate (and Not Regulate) Social Media*…, p. 75.

1.2. Do otimismo à realidade: o problema das *fake news*

A criação de um ambiente aberto às discussões éticas, culturais, políticas e de tantas outras naturezas gerou, em um primeiro momento, entusiasmo, levando-se a cogitar que a internet e suas novas tecnologias representariam uma renovação da esperança de realização da democracia.[17]

A possibilidade de qualquer pessoa produzir e publicar conteúdo autoral ou de terceiros em uma escala global, por meio das redes sociais, revelou o potencial da internet para criar uma comunidade democrática global e viabilizar expressivas mobilizações contra governos autocráticos. Exemplo didático disso foram os movimentos pró-democracia da Primavera Árabe nos idos de 2010.[18]

Não demorou muito para que o otimismo desse lugar a uma visão mais realista e, até mesmo, pessimista sobre as redes sociais. Ao invés de a democracia digital estimular o debate público, observou-se com maior frequência a proliferação de manifestações antidemocráticas, discursos de ódio e conteúdos danosos que abriram espaço para o radicalismo e extremismo.[19_20]

---

[17] SCHREIBER, Anderson. Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro. In: LUCCA, Newton de; SIMÃO FILHO, Adalberto; LIMA, Cíntia Rosa Pereira. *Direito e Internet III: Marco Civil da Internet, Lei nº 12.965/2014*, Tomo II. São Paulo: Quartier Latin, 2015, pp. 277-305.

[18] 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. 26.

[19] SCHREIBER, Anderson. *Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro...*, pp. 277-305.

[20] Há certo consenso de que, embora não seja a causa primária ou única do radicalismo, a internet seria uma facilitadora e catalizadora de manifestações de atos políticos violentos (KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* (December 16, 2019). Forthcoming, N. Persily & J. Tucker, Social Media and Democracy: The State of the Field and Prospects for Reform (Cambridge University Press), p. 36. Disponível em: <https://ssrn.com/abstract=3504930>. Acesso em 29 nov. 2024).

As redes sociais se tornaram terreno fértil para a propagação de notícias falsas, conhecidas como *fake news*.[21–22] Embora não seja um fenômeno novo,[23] a disseminação de *fake news* tomou, com o alcance das redes sociais, proporções jamais enfrentadas, ganhando contornos social, político e juridicamente relevantes.

O referendo do Brexit no Reino Unido e as eleições presidenciais americanas, em 2016, são elencados como os primeiros eventos catalizadores de expressiva propagação de *fake news* nas redes sociais. Esses eventos causaram preocupações à proteção do processo democrático e dos direitos fundamentais, diante de *(i)* campanhas massivas de desinformação, inclusive com a participação de representantes eleitos, candidatos e/ou governantes de outros países, *(ii)* microdirecionamento de propaganda com potencial de alteração de resultados eleitorais e *(iii)* ataques antidemocráticos, discursos de ódio e conteúdos ilícitos.[24]

A recente eleição presidencial de 2024 nos Estados Unidos da América ("EUA") foi marcada pela propagação sem precedentes de desinformação nas redes sociais, inclusive por presidenciáveis e relevantes personalidades públicas.[25] Passados oito anos dos primeiros eventos de escala global de propagação de

---

[21] Hunt Allcott e Matthew Gentzkow definem *fake news* como "notícias que são intencional e comprovadamente falsos, e podem levar os leitores ao engano". (ALLCOTT, Hunt; GENTZKOW, Matthew (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, vol. 31, n. 2, 2017, pp. 211–236, p. 213. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211>. Acesso em 29 nov. 2024). No entanto, há uma certa limitação do termo "fake news" em descrever os fenômenos complexos e variados de desinformação e informações equivocadas. Além disso, a expressão tem sido apropriada indevidamente por grupos políticos para atacar a liberdade de imprensa e propagar desinformação. Nesse sentido, para informações sobre as críticas à utilização do termo "fake news", ver: WARDLE, Claire; DERAKHSHAN, Hossein (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, DGI (2017)09, p. 5. Disponível em: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>. Acesso em 29 nov. 2024; e TOFFOLI, Dias. Fake News: desinformação e liberdade de expressão. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2022, p. 33.

[22] SCHREIBER, Anderson. *Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro*..., p. 5.

[23] Um exemplo histórico seria o "Grande Engano da Lua", no qual o New York Sun publicou, no ano de 1835, "uma série de artigos sobre a descoberta de vida na lua." (ALLCOTT, Hunt; GENTZKOW, Matthew. *Social media and fake news in the 2016 election*..., pp. 213-214).

[24] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 26.

[25] LYNGAAS, Sean. *Eleição dos EUA 'vê quantidade sem igual de desinformação', diz chefe de segurança cibernética*. CNN, 4 nov. 2024. Disponível em: <https://www.cnnbrasil.com.br/internacional/eleicoes-nos-eua-2024/eleicao-dos-eua-ve-quantidade-sem-igual-de-desinformacao-diz-chefe-de-seguranca-cibernetica/>. Acesso em 29 nov. 2024.

desinformação, as *fake news* ainda são um dos maiores desafios enfrentados pela sociedade contemporânea.

No cenário brasileiro, as redes sociais também foram massivamente utilizadas para a disseminação de notícias falsas durante as eleições presidenciais dos anos de 2018 e 2022. O enfrentamento da desinformação, em 2022, exigiu dos órgãos de controle e fiscalização, capitaneados pela Justiça Eleitoral, o combate às *fake news* a fim de preservar a integridade do processo eleitoral.[26]

O fenômeno não se limita às questões preponderantemente jurídicas. As *fakes news* vêm sendo apontadas ainda como um grande problema de saúde pública.[27]-[28] O avanço das notícias falsas em matéria de saúde pública contribuiu para o renascimento do famigerado movimento antivacina.[29] A cobertura de vacinação de brasileiros contra sarampo, caxumba e rubéola (Tríplice Viral D1) despencou de 93,1% para 71,49%, entre 2019 e 2021.[30] Na prática, houve o aumento do risco de morte e deficiências física e mental graves a milhares de crianças e adolescentes no país.[31] A campanha de divulgação na internet de notícias falsas relacionadas à saúde e à vacinação se intensificou, atingindo proporções ainda maiores durante a pandemia da COVID-19.[32]

---

[26] O Tribunal Superior Eleitoral aprovou, em 20 de outubro de 2022, a Resolução 23.714, que dispõe sobre o enfrentamento da desinformação que compromete a integridade do processo eleitoral. A Resolução 23.714/2022 veda, no artigo 2º, "a divulgação ou compartilhamento de fatos sabidamente inverídicos ou gravemente descontextualizados que atinjam a integridade do processo eleitoral, inclusive os processos de votação, apuração e totalização de votos". (BRASIL. Tribunal Superior Eleitoral. Disponível em: <https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022>. Acesso em 9 dez. 2024.

[27] ANJOS, Alise Silva Martins; CASAM, Priscila Carla; MAIA, Janize Silva. *As fake news e seus impactos na saúde da sociedade*. Pub Saúde, 5, a141, 2021, p. 1.

[28] À guisa de exemplo, o 5º Relatório da Segurança Digital, elaborado pela PSafe, no segundo semestre de 2018, já afirmava que 41,6% das notícias falsas abordaram o tema da saúde, com especial enfoque às campanhas de vacinação. (PSAFE. Relatório da segurança digital no Brasil: terceiro trimestre – 2018. Disponível em <https://www.psafe.com/dfndr-lab/pt-br/relatorio-da-seguranca-digital>. Acesso em 30 nov. 2024).

[29] GRAGNANI, Juliana; SENRA, Ricardo. BBC News. Movimento antivacina é criminoso, diz Drauzio Varella. Publicado em 26 jun. 2019. Disponível em: <https://www.bbc.com/portuguese/geral-48780905>. Acesso em 9 dez. 2024.

[30] UNICEF. *3 em cada 10 crianças no Brasil não receberam vacinas que salvam vidas, alerta UNICEF*. Publicado em 27 abr. 2022. Disponível em: <https://www.unicef.org/brazil/comunicados-de-imprensa/3-em-cada-10-criancas-no-brasil-nao-receberam-vacinas-que-salvam-vidas>. Acesso em 29 nov. 2024.

[31] BUTANTAN. *Queda nas taxas de vacinação no Brasil ameaça a saúde das crianças.* São Paulo. Publicado em 7 mar. 2022. Disponível em: <https://butantan.gov.br/noticias/queda-nas-taxas-de-vacinacao-no-brasil-ameaca-a-saude-das-criancas>. Acesso em 29 nov. 2024.

[32] MARTIN, Ana.. *Vacina contra Covid no PNI: 'Fato ou Fake' desmentiu dezenas de informações falsas; veja 10 delas*. São Carlos e Região. Portal G1, publicado em 19 jul. 2023. Disponível em: <https://g1.globo.com/sp/sao-carlos-regiao/noticia/2023/07/19/vacina-contra-covid-no-pni-fato-ou-fake-desmentiu-dezenas-de-informacoes-falsas-veja-10-delas.ghtml>. Acesso em 9 dez. 2024.

Para enfrentamento do problema das *fake news*, Carlos Affonso Souza e Chiara de Teffé defendem a adoção de iniciativas ligadas à educação digital e informacional dos cidadãos para formação de uma consciência mais crítica. Os autores indicam dez recomendações, elaboradas por instituições especializadas, para que as pessoas possam identificar conteúdos falsos ou manipulados, que envolvem cautelas como, por exemplo, a prudência na checagem do conteúdo, da credibilidade da fonte e outras ações como a pesquisa em agências de *fact-checking*, que avaliam a autenticidade dos fatos.[33]

De fato, essas questões demandam atenção especial da literatura. A esfera pública digital criada pelas redes sociais é moldada notadamente por normas privadas, através da moderação de conteúdo que se baseia nos termos de uso de plataformas transnacionais.[34] Além de descaracterizar em certa medida a percepção de descentralização do espaço público virtual, o estado da arte aponta para preocupações em torno da governança privada do discurso público *online*.

1.3. A esfera pública digital: a governança privada do discurso público

No modelo de governança privada, as condições práticas do discurso público fluem por uma elaborada infraestrutura de comunicação privada – isto é, as plataformas digitais –, influenciando quem o controla, limita e censura. Em outras palavras, a liberdade de expressão está sujeita às decisões dos entes privados que detêm essa infraestrutura de comunicação.[35]

As grandes plataformas transnacionais que detêm as redes sociais com o maior número de usuários mensais ativos pelo mundo, como *Facebook* e *Instagram* (*Meta*), *Youtube* (*Google*) e *X*, antigo *Twitter* (X Corp.), situam-se entre os Estados Nacionais e os indivíduos na curadoria *online* do discurso público.[36] Essas

---

[33] SOUZA, Carlos Affonso; TEFFÉ, Chiara Spadaccini. Fake news e eleições: identificado e combatendo a desordem informacional. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., p. 309.
[34] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 26.
[35] BALKIN, Jack M., *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*..., p. 4.
[36] *Ibidem*, p. 1.

plataformas operam com os novos governantes (*New Governors*) do discurso público *online*, sendo parte central desse sistema "triangular".[37]

Em termos práticos, essa mudança resulta na transferência de parte relevante, embora não completa, do controle da liberdade de expressão na esfera pública digital e da ponderação do seu alcance face à necessária proteção dos direitos fundamentais, até então exercidos pelos Estados Nacionais, para essas plataformas transnacionais.

Esse poder das plataformas digitais, que permeia a atual discussão sobre a necessidade de reequilíbrio da equação entre liberdade econômica e a garantia de direitos fundamentais à liberdade de expressão, tem sido cunhado com a expressão "Sociedade das Plataformas", que considera justamente o protagonismo das plataformas privadas enquanto relevantes atores econômicos e políticos.[38]

O protagonismo das redes sociais e a evolução do modelo de governança privada do discurso público têm despertado movimentos de pressão de governos, cidadãos e usuários finais para responder ao problema das *fake news*,[39] a fim de se construir uma esfera pública digital saudável e livre dos abusos cometidos à pretexto de uma noção de liberdade de expressão absoluta.

Para resolver esse problema, uma variedade de atores públicos e privados tem instigado as plataformas digitais a desenvolver seus programas, algoritmos e procedimentos para controle do discurso, além de ajudar os usuários finais a tomar decisões sobre quais tipos de notícias/conteúdo devem ler e confiar. Trata-se de um novo sistema de regulação do discurso público que deve estimular o desenvolvimento de uma regulação capaz de assegurar uma internet livre e plural.[40]

O debate sobre a regulação das redes sociais deve ser conduzido com o objetivo de promover os valores associados à liberdade de expressão, assim como a responsabilidade das redes sociais pela proteção de uma esfera pública digital saudável. Esses pontos serão detalhados a seguir no subcapítulo 1.6 *infra*, que abordará as razões pelas quais as redes sociais devem ser reguladas.

---

[37] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.603.
[38] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil. *Cadernos Adenauer*, XXIV, 2023, nº 1, p. 63-87, p. 69.
[39] BALKIN, Jack M., *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*…, p. 66.
[40] *Ibidem*, pp. 66-67.

Antes de tratar desse tema, passa-se a se debruçar sobre a liberdade de comunicação como pressuposto à realização da opinião pública, assim como a evolução da legislação brasileira e aspectos doutrinários e jurisprudenciais a respeito do direito à liberdade de expressão.

1.4. A concepção contemporânea da liberdade de expressão

Na sua concepção contemporânea, a liberdade de expressão pode ser compreendida, em sentido amplo, como o conjunto de direitos que decorrem da liberdade de comunicação, direitos esses que englobam a liberdade de expressão em sentido estrito (manifestação de pensamento e de opinião), a liberdade de criação e de impressa e o direito de informação.[41]

Partindo do pressuposto de que a concepção de liberdade de expressão deve ser a mais ampla possível, desde que resguardada a operacionalidade do direito, decorrem dessa noção outros direitos relacionados à liberdade de expressão como o direito de informar e de ser informado, o direito de resposta, o direito de réplica política, a liberdade de reunião, a liberdade religiosa etc.[42]

A liberdade de comunicação está relacionada, por sua vez, ao conjunto de direitos que tutelam a manifestação de pensamento e informação.[43] O exercício desses direitos tem ligação direta com a concretização do debate e a formação da

---

[41] TÔRRES, Fernanda Carolina. *O direito fundamental à liberdade de expressão e sua extensão*. Revisão de Informação Legislativa. Senado Federal. Ano 50 Número 200 out./dez. 2013, p. 62. Disponível em: <https://www12.senado.leg.br/ril/edicoes/50/200/ril_v50_n200_p61.pdf/>. Acesso em 27 dez. 2024.

[42] *Ibidem*, p. 63.

[43] SILVA, José Afonso da. *Curso de direito constitucional positivo.* 22ª ed. São Paulo: Malheiros, 2003, p. 242: "A liberdade de comunicação consiste num conjunto de direitos, formas, processos e veículos, que possibilitam a coordenação desembaraçada da criação, expressão e difusão do pensamento e da informação. É o que se extrai dos incisos IV, V, IX, XII e XIV do art. 5º combinados com os arts. 220 a 224 da Constituição. Compreende ela as formas de criação, expressão e manifestação do pensamento e da informação, e a organização dos meios de comunicação (...)."

opinião pública,[44] com o objetivo de assegurar a manifestação de pensamento que critique valores e ideias sedimentados no tecido social.[45]

Colocada a liberdade de expressão nesses termos, surge o reconhecimento de que existe uma contínua tensão em torno da circulação de ideias, existindo agentes com capacidade política, financeira e/ou técnico-jurídica para manipular ou impedir a manifestação de pensamento. Desse modo, há uma relação de assimetria entre aquele que se expressa e seus potenciais censores.[46]

1.4.1. A liberdade de comunicação como pressuposto à realização da opinião pública

A condução da opinião pública possui uma dimensão inafastável do poder, cuja função é definir a legitimidade das *verdades* e *dúvidas*. Por esse motivo, a tutela da liberdade de expressão previne a constituição de discursos que beneficiam os governantes – confundidos, de forma equivocada, única e exclusivamente com a figura do Estado – em detrimento da sociedade.[47]

---

[44] NOELLE-NEUMANN, Elisabeth. *A espiral do silêncio – Opinião pública: nosso tecido social*. Cristian Derosa (Trad.). Florianópolis: Estudos Nacionais, 2017, pp. 304-305: "A opinião pública como processo racional fixa-se especialmente na participação democrática e no intercâmbio de pontos de vista diferentes sobre os assuntos públicos, assim como na exigência de que o governo leve em conta estas idéias e a preocupação com a possibilidade de manipulação da opinião pública pelo poder do estado e do capital, através dos meios de comunicação e a técnica moderna (Habermas, 1962). A opinião pública como controle social busca garantir um nível suficiente de consenso social sobre os valores e os objetivos comuns. Segundo esse conceito, o poder da opinião pública é tão grande que não pode ser ignorado pelo governo e tampouco pelos membros individuais da sociedade. Tal poder procede da ameaça de isolamento que a sociedade dirige contra os indivíduos e os governos desviados, e do medo do isolamento devido a natureza social do homem."

[45] SARMENTO, Daniel. Comentários ao art. 5.º, incisos IV, V e IX. In: CANOTILHO, J. J. Gomes; MENDES, Gilmar Ferreira; SARLET, Ingo Wolfgang, STRECK, Lenio Luiz (coord.). *Comentários à Constituição do Brasil*, São Paulo: Saraiva/Almedina, 2013, *ebook*, p. 555: "A proibição da censura é um dos aspectos centrais da liberdade de expressão. É natural a inclinação dos regimes autoritários em censurar a difusão de ideias e informações que não convêm aos governantes. Mas, mesmo fora das ditaduras, a sociedade muitas vezes reage contraposições que questionem os seus valores mais encarecidos e sedimentados, e daí pode surgir a pretensão das maiorias de silenciar os dissidentes."

[46] MENDES, Gilmar Ferreira; BRANCO, Paulo Gustavo Gonet. *Curso de Direito Constitucional*. 9ª ed. São Paulo: Saraiva, 2014, pp. 264-265: "A liberdade de expressão, enquanto direito fundamental, tem, sobretudo, um caráter de pretensão a que o Estado não exerça censura. Não é o Estado que deve estabelecer quais opiniões que merecem ser tidas como válidas e aceitáveis; essa tarefa cabe, antes, ao público a que essas manifestações se dirigem. Daí a garantia do art. 220 da Constituição brasileira. Estamos, portanto, diante de um direito de índole marcadamente defensiva – direito a uma abstenção pelo Estado de uma conduta que interfira sobre a esfera da liberdade do indivíduo."

[47] BOBBIO, Noberto; MATTEUCI, Nicola; PASQUINO, Gianfranco. *Dicionário político*. Vol. 1. 11ª ed. Brasília: Universidade de Brasília, 1998. (ebook): "Opinião Pública. (...) A existência da opinião pública é um fenômeno da época moderna: pressupõe uma sociedade civil distinta do Estado, uma sociedade livre e articulada, onde existam centros que permitam a formação de opiniões não

Paulo Bonavides explica que os direitos da liberdade "têm por titular o indivíduo, são oponíveis ao Estado, traduzem-se como faculdade ou atributos da pessoa e ostentam uma subjetividade que é seu traço característico; enfim, são direitos de resistência ou de oposição perante o Estado".[48]

Em exame da liberdade de expressão da Lei Fundamental alemã, no entanto, Konrad Hesse afirmou que a liberdade de expressão é mais do que um direito subjetivo de defesa; é um direito de cooperação política. Ao vedar a prática de censura, o Texto Constitucional torna a liberdade de expressão uma *garantia institucional* necessária ao funcionamento da opinião pública e, por consequência, do próprio Estado Democrático de Direito, assegurando o pluralismo político.[49]

No Brasil, o artigo 5°, incisos IV e IX, e o artigo 220, *caput*, §§ 1° e 2°, da Constituição Federal, tutelam a liberdade de expressão, assegurando a possibilidade de manifestar opiniões, ideias e pensamentos, sem temor de repressão, e resguardando também o direito à informação e à comunicação. Para implementar essa garantia institucional, Gilmar Mendes ensina que o legislador deve restringir a liberdade de expressão desde que a medida vise garantir e efetivar essa liberdade.[50]

---

individuais, como jornais e revistas, clubes e salões, partidos e associações, bolsa e mercado, ou seja, um público de indivíduos associados, interessado em controlar a política do Governo, mesmo que não desenvolva uma atividade política imediata. Por isso, a história do conceito de opinião pública coincide com a formação do Estado moderno que, com o monopólio do poder, privou a sociedade corporativa de todo o caráter político, relegando o indivíduo para a esfera privada da moral, enquanto a esfera pública ou política foi inteiramente ocupada pelo Estado. Mas, após o advento da burguesia, ao constituir-se dentro do Estado uma sociedade civil dinâmica e articulada, foi se formando um público que não quer deixar, sem controle, a gestão dos interesses públicos na mão dos políticos. A opinião pública foi levada deste modo a combater o conceito de segredo de Estado, a guarda dos *arcaria imperii* e a censura, para obter o máximo de 'publicidade' dos atos do Governo."

[48] BONAVIDES, Paulo. *Curso de Direito Constitucional*. 15ª ed. São Paulo: Malheiros, 2004, p. 563-564.

[49] HESSE, Konrad. *Elementos de Direito Constitucional da Alemanha Federal*. Luís Afonso Heck (tradutor). Porto Alegre: S.A. Fabris, 1998, pp. 302-303: "O alcance completo dessas garantias abre-se, também aqui, somente com vista ao seu caráter duplo: elas são, por um lado, direitos subjetivos, e, precisamente, tanto no sentido de direitos de defesa como no de direitos de cooperação política; por outro, elas são prescrições de competência negativa e elementos constitutivos da ordem objetiva democrática e estatal-jurídica. Sem a liberdade de manifestação da opinião e liberdade de informação, sem um espaço público de comunicação política não pode nascer, o desenvolvimento de iniciativas e alternativas pluralistas, assim como a 'formação preliminar da vontade política' não são possíveis. A publicidade da vida política e a participação política em um processo livre e aberto não se podem desenvolver. Liberdade de opinião é, por causa disso, para a ordem democrática da Lei Fundamental, 'simplesmente constitutiva'."

[50] MENDES, Gilmar. *Direitos fundamentais e controle de constitucionalidade*. 4ª ed. São Paulo: Saraiva, 2012, p. 645: "(...) pode ser observado nos textos constitucionais que, como o art. 220 da Constituição brasileira de 1988, contêm cláusula proibitiva de qualquer restrição às liberdades de expressão e de imprensa. Ao mesmo tempo em que prescrevem a não restrição dessas liberdades, tais textos não apenas permitem, como também obrigam a intervenção legislativa no sentido de sua

O argumento é compartilhado por parte da doutrina, que, identificando um deslocamento do debate público às redes sociais contemporâneas, compreende que existem nelas, mais do que em qualquer outro espaço, o exercício abusivo do direito de liberdade de expressão, que é maximizado por uma quantidade de participantes sem precedentes na história.[51]

O exercício da liberdade não se concretizou nas redes sociais de forma "natural", mas como resultado da cultura jurídica que permeia as redes sociais. Rodrigo Vidal Nitrini afirma que foi questão de tempo para que a liberdade de expressão exercida na internet violasse bens jurídicos caros a determinados ordenamentos jurídicos, como também ocorre no contexto brasileiro. Isso porque, as empresas de tecnologia e comunicação mais relevantes do mundo operam a partir do ordenamento jurídico estadunidense,[52] marcadamente libertário no exercício da liberdade de expressão.[53]

1.4.2. Evolução legislativa da liberdade de comunicação no Brasil

Nas primeiras formulações constitucionais, o direito de comunicação era previsto sem dependência de censura ou licença do Poder Público, sendo o

promoção e efetividade. Entre concepções liberais, individuais ou subjetivas, por um lado, e outras concepções cívicas, republicanas, democráticas ou objetivas, por outro, o aparente paradoxo das liberdades de expressão, de informação e de imprensa tem sido enfrentado pelas Cortes Constitucionais com base em um postulado que hoje transparece quase como uma obviedade: as restrições legislativas são permitidas e até exigidas constitucionalmente quando têm o propósito de proteger, garantir e efetivar tais liberdades."

[51] SCHREIBER, Anderson. *Marco Civil da Internet: Avanço ou Retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro*..., p. 6: "O único caminho, portanto, é a aplicação de normas que assegurem que a liberdade de expressão não seja exercida em desfavor de si própria. Como já se disse no passado em relação à liberdade de contratar, também a liberdade de expressão é "autofágica", no sentido de que, em qualquer ambiente em que haja desigualdade de forças, a liberdade de expressão do mais forte tende a subjugar a liberdade de expressão do mais fraco. Em cenários desiguais, a ausência de normas não costuma resultar em maior liberdade, mas, ao contrário, em *mera aparência de liberdade*, na medida em que a omissão normativa beneficia tão-somente aqueles que, detendo maior poderio econômico e técnico, se veem, finalmente, livres para perseguir seus interesses sem precisar respeitar regras instituídas no interesse da sociedade como um todo."

[52] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., pp. 35-36: "(...) três empresas americanas – Youtube, Facebook e Twitter – estabeleceram-se como as plataformas dominantes para compartilhamento global de conteúdos. Nesse contexto, o direito americano assumiu uma importância fundadora para o desenvolvimento da internet comercial e suas consequências para as formas de exercício de – ou de restrições a – direitos fundamentais no mundo digital, nos mais diversos países. No caso da liberdade de expressão, isso significa que as regras legais e a cultura jurídica daquele país moldaram, em larga medida, a governança privada de discursos pelas grandes redes sociais."

[53] Para mais detalhes sobre a evolução da jurisprudência da Suprema Corte Estadunidense para uma posição libertária, remeta-se *a* FISS, Owen M. *The irony of free speech*. Cambridge: Harvard University Press, 1998, p. 79 e ss.

anonimato vedado e assegurada a responsabilidade pelo exercício abusivo desse direito (artigo 179, IV, da Carta Imperial de 1824; artigo 72, § 12, da Constituição de 1891; e o artigo 113, n. 9, da Constituição de 1934).[54]

Nessas experiências, leis ordinárias indicavam uma extensa lista de matérias que eram consideradas abusivas, como, por exemplo: (i) "doutrinas dirigidas a destruir as verdades fundamentaes da existencia de Deus, e da immortalidade da Alma" (artigo 2º, n. 4º, Lei de 20 de setembro de 1830); e (ii) "offensa à moral publica ou aos bons costumes" (artigo 5º do Decreto 4.743/1923). Contudo, em exemplo interessante, constava como exercício abusivo da liberdade de expressão a comunicação de "notícias falsas, ou noticiar fatos verdadeiros, umas e outros, porém, tendenciosamente, por forma a provocar alarme social, ou perturbação da ordem pública" (artigo 11 do Decreto 24.776/1934).

Com a promulgação da Constituição de 1937, já no contexto do Estado Novo, ficou determinado que o legislador poderia prescrever a censura prévia para garantir a paz, ordem, bons costumes, moralidade e segurança públicas (artigo 122, n. 15, alíneas *a*, *b* e *c*).[55] Após a queda da ditadura varguista, esse direito passa por um *continuum* que impõe censura prévia aos espetáculos e diversões públicas

---

[54] BRASIL. *Constituição política do Imperio do Brazil (de 25 de março de 1824)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao24.htm. Acesso em 07 mar. 2025: "Art. 179 (...) IV. Todos podem communicar os seus pensamentos, por palavras, escriptos, e publica-los pela Imprensa, sem dependencia de censura; com tanto que hajam de responder pelos abusos, que commetterem no exercicio deste Direito, nos casos, e pela fórma, que a Lei determinar".
BRASIL. *Constituição da República dos Estados Unidos do Brasil (24 de fevereiro de 1891)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao91.htm. Acesso em 07 mar. 2025: "Art. 72 (...) § 12. Em qualquer assumpto é livre a manifestação do pensamento pela imprensa ou pela tribuna, sem dependencia de censura, respondendo cada um pelos abusos que commetter, nos casos e pela fórma que a lei determinar. Não é permittido o anonymato."
BRASIL. *Constituição da República dos Estados Unidos do Brasi (16 de julho de 1934)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao34.htm. Acesso em 07 mar. 2025: "Art. 113 (...) 9) Em qualquer assumpto é livre a manifestação do pensamento, sem dependencia de censura, salvo quanto a espectaculos e diversões publicas, respondendo cada um pelos abusos que commetter, nos casos e pela fórma que a lei determinar. Não é permittido o anonymato. É assegurado o direito de resposta. A publicação de livros e periodicos independe de licença do poder publico. Não será, porém, tolerada propaganda de guerra ou de processos violentos para subverter a ordem politica ou social."

[55] BRASIL. *Constituição dos Estados Unidos do Brasil (10 de novembro de 1937)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao37.htm. Acesso em 07 mar. 2025: "Art. 112 (...) 15) todo cidadão tem o direito de manifestar o seu pensamento, oralmente, ou por escrito, impresso ou por imagens, mediante as condições e nos limites prescritos em lei. A lei pode prescrever: a) com o fim de garantir a paz, a ordem e a segurança pública, a censura prévia da imprensa, do teatro, do cinematógrafo, da radiodifusão, facultando à autoridade competente proibir a circulação, a difusão ou a representação; b) medidas para impedir as manifestações contrárias à moralidade pública e aos bons costumes, assim como as especialmente destinadas à proteção da infância e da juventude; c) providências destinadas à proteção do interesse público, bem-estar do povo e segurança do Estado."

(artigo 141, §5º, da Constituição de 1946).[56] À guisa de exemplo, no âmbito infraconstitucional, a lei que regulava a liberdade de expressão e de imprensa também entendia como abusivo "publicar notícias falsas ou divulgar fatos verdadeiros, truncados ou deturpados, que provoquem alarma social ou perturbação da ordem pública" (artigo 9º, alínea *b*, da Lei 2.083/1953).

Com a promulgação da Constituição de 1967, houve a reprodução da mesma fórmula constitucional para limitar o exercício da liberdade de expressão (artigo 150, § 8º)[57], cabendo ao legislador ordinário reproduzir as proibições já aventadas nas últimas experiências constitucionais, sendo possível censurar jornais, revistas, periódicos, empresas de notícia e imprensa durante a decretação de estado de sítio (artigo 1º, §2º, da Lei 5.250/1967 – Lei da Imprensa).[58] Seguiu a mesma linha de regulação da liberdade de expressão o artigo 153, §3º, da Emenda Constitucional nº

---

[56] BRASIL. *Constituição dos Estados Unidos do Brasil (18 de setembro de 1946)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao46.htm. Acesso em 07 mar. 2025: "Art. 141 - (...) § 5º - É livre a manifestação do pensamento, sem que dependa de censura, salvo quanto a espetáculos e diversões públicas, respondendo cada um, nos casos e na forma que a lei preceituar pelos abusos que cometer. Não é permitido o anonimato. É assegurado o direito de resposta. A publicação de livros e periódicos não dependerá de licença do Poder Público. Não será, porém, tolerada propaganda de guerra, de processos violentos para subverter a ordem política e social, ou de preconceitos de raça ou de classe."

[57] [57] BRASIL. *Constituição da República Federativa do Brasil (24 de janeiro de 1967)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao67.htm. Acesso em 07 mar. 2025: "Art. 150 - (...) § 8º - É livre a manifestação de pensamento, de convicção política ou filosófica e a prestação de informação sem sujeição à censura, salvo quanto a espetáculos de diversões públicas, respondendo cada um, nos termos da lei, pelos abusos que cometer. É assegurado o direito de resposta. A publicação de livros, jornais e periódicos independe de licença da autoridade. Não será, porém, tolerada a propaganda de guerra, de subversão da ordem ou de preconceitos de raça ou de classe.".

[58] BRASIL. *Lei 5.250 (9 de fevereiro de 1967)*. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l5250.htm. Acesso em 07 mar. 2025: "Art. 1º - (...) § 2º O disposto neste artigo não se aplica a espetáculos e diversões públicas, que ficarão sujeitos à censura, na forma da lei, nem na vigência do estado de sítio, quando o Govêrno poderá exercer a censura sôbre os jornais ou periódicos e emprêsas de radiodifusão e agências noticiosas nas matérias atinentes aos motivos que o determinaram, como também em relação aos executores daquela medida".

1/1969,[59] mantendo a vigência da Lei de Imprensa até a declaração de sua não recepção pela Constituição de 1988.[60]

Os ventos mudam com o fim da Ditadura Civil-Militar, compreendida entre 1964 e 1988,[61] e a promulgação da Constituição Federal de 1988, que concedeu significativa ênfase à tutela da liberdade de expressão. No artigo 5º, IV e V, da Carta Constitucional, consagrou-se que "é livre a manifestação do pensamento, sendo vedado o anonimato", concedendo direito de resposta proporcional ao agravo, assim como a indenização por danos materiais ou morais.

Contrário à censura prévia a espetáculos públicos, o inciso IX do artigo 5° reforça que é livre a expressão da atividade intelectual, artística, científica e de comunicação, independentemente de censura ou licença, rompendo com o autoritarismo imposto desde a Constituição de 1937. Por fim, o inciso XIV protege o direito à informação, garantindo o sigilo da fonte quando necessário ao exercício profissional e, de modo complementar, o artigo 220 da Constituição reafirma a liberdade de manifestação na comunicação social, vedando qualquer forma de censura de natureza política, ideológica ou artística.

No âmbito internacional, a liberdade de expressão também encontra guarida em tratados que o Brasil incorporou ao direito interno. O artigo 13, n. 1 a 5, da Convenção Americana de Direitos Humanos (Pacto de San José da Costa Rica) estabelece que toda pessoa tem o direito à liberdade de pensamento e de

---

[59] BRASIL. *Emenda Constitucional nº 1 (17 de outubro de 1969)*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc_anterior1988/emc01-69.htm. Acesso em 07 mar. 2025: "Art. 153 – (...) § 8º É livre a manifestação de pensamento, de convicção política ou filosófica, bem como a prestação de informação independentemente de censura, salvo quanto a diversões e espetáculos públicos, respondendo cada um, nos têrmos da lei, pelos abusos que cometer. É assegurado o direito de resposta. A publicação de livros, jornais e periódicos não depende de licença da autoridade. Não serão, porém, toleradas a propaganda de guerra, de subversão da ordem ou de preconceitos de religião, de raça ou de classe, e as publicações e exteriorizações contrárias à moral e aos bons costumes."

[60] No julgamento da ADPF 130 foi declarada a incompatibilidade da Lei de Imprensa (Lei 5.250/1967) com a Constituição de 1988, por entender que seus dispositivos violavam a liberdade de comunicação. A Corte reafirmou que a censura prévia é vedada no regime democrático brasileiro, destacando a importância da liberdade de manifestação para a consolidação da democracia. O STF manteve o direito de resposta e de reparação por danos, desde que respeitados os limites constitucionais, reconhecendo o papel da imprensa na promoção da transparência e *accountability*. Essa decisão foi um marco na proteção da liberdade de expressão no Brasil (BRASIL. Supremo Tribunal Federal, Tribunal Pleno, ADPF 130, Rel. Min. Carlos Ayres Brito, j. 30 abr. 2009).

[61] BRASIL. Câmara dos Deputados. *50 anos do Golpe de 1964*. <https://www2.camara.leg.br/atividade-legislativa/plenario/discursos/escrevendohistoria/destaque-de-materias/golpe-de-1964>. Acesso em 07 mar. 2025.

expressão.[62] De forma similar, o artigo 19, n. 2 e 3, alíneas *a* e *b*, do Pacto Internacional dos Direitos Civis e Políticos, ratificado pelo Brasil em 1992, reforça esses preceitos, consolidando o compromisso do país com os princípios de liberdade e pluralidade de expressão.[63]

Com a ampliação do acesso à internet, a construção do debate público foi alterado significativamente com a participação direta e constante do público na circulação de ideias. Nesse contexto, o legislador brasileiro editou o Marco Civil da Internet (Lei 12.965/2014) para assegurar a liberdade de expressão aos usuários, com vedação à censura e responsabilidade subsidiária do provedor de internet, em caso de descumprimento judicial (artigo 19, *caput*).[64]

Em linha com a tradição jurídica, o uso da internet no Brasil, regulado pelo Marco Civil da Internet, também se vinculou à promoção de direitos humanos, direitos da personalidade e a pluralidade e diversidade social, restando presente os

---

[62] BRASIL. *Decreto nº 678 – Pacto de San José da Costa Rica (6 de novembro de 1992)*. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto/d0678.htm. Acesso em 07 mar. 2025: "Artigo 13 – Liberdade de Pensamento e de Expressão – 1. Toda pessoa tem direito à liberdade de pensamento e de expressão. Esse direito compreende a liberdade de buscar, receber e difundir informações e idéias 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. 2. O exercício do direito previsto no inciso precedente não pode estar sujeito a censura prévia, mas a responsabilidades ulteriores, que devem ser expressamente fixadas pela lei a ser necessárias para assegurar: a) o respeito aos direitos ou à reputação das demais pessoas; ou b) a proteção da segurança nacional, da ordem pública, ou da saúde ou da moral públicas. 3. Não se pode restringir o direito de expressão por vias ou meios indiretos, tais como o abuso de controles oficiais ou particulares de papel de imprensa, de freqüências radioelétricas ou de equipamentos e aparelhos usados na difusão de informação, nem por quaisquer outros meios destinados a obstar a comunicação e a circulação de idéias e opiniões. 4. A lei pode submeter os espetáculos públicos a censura prévia, com o objetivo exclusivo de regular o acesso a eles, para proteção moral da infância e da adolescência, sem prejuízo do disposto no inciso 2. 5. A lei deve proibir toda propaganda a favor da guerra, bem como toda apologia ao ódio nacional, racial ou religioso que constitua incitação à discriminação, à hostilidade, ao crime ou à violência."

[63] BRASIL. *Decreto 592 – Pacto Internacional dos Direitos Civis e Políticos* (6 de julho de 1992). Disponível em: https://www.planalto.gov.br/ccivil_03/decreto/1990-1994/d0592.htm. Acesso em 07 mar. 2025: "Artigo 19 (...) 2. Toda pessoa terá direito à liberdade de expressão; esse direito incluirá a liberdade de procurar, receber e difundir informações e idéias 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. 3. O exercício do direito previsto no parágrafo 2 do presente artigo implicará deveres e responsabilidades especiais. Conseqüentemente, poderá estar sujeito a certas restrições, que devem, entretanto, ser expressamente previstas em lei e que se façam necessárias para: a) assegurar o respeito dos direitos e da reputação das demais pessoas; b) proteger a segurança nacional, a ordem, a saúde ou a moral públicas."

[64] BRASIL. *Lei nº 12.695 – Marco Civil da Internet (23 de abril de 2014*). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em 07 mar. 2025: "Art. 19. 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."

elementos que possibilitariam uma restrição da liberdade de expressão, caso subsista fatos que indiquem sua disfuncionalidade.[65]

Com base nas normas constitucionais, infraconstitucionais e principais tratados internacionais, é pacífico que, como regra, a *censura prévia*[66] é vedada no ordenamento jurídico brasileiro.[67] O exercício da liberdade de expressão deve ser submetido ao controle *a posteriori* da sociedade.[68] Também é possível afirmar, com base nos referidos tratados, que existe um sintético, porém direcionado, regime que possibilita restringir a liberdade de expressão, caso seu exercício viole direitos de terceiros[69] ou prejudique a segurança nacional ou a ordem e saúde públicas. Assim, a liberdade de expressão só é tutelada pelo ordenamento jurídico quando exercida em consonância com as demais normas e princípios constitucionais.[70]

O ministro Dias Toffoli afirma que, embora o Supremo Tribunal Federal – STF tenha construído sólida jurisprudência em defesa da liberdade de expressão,[71]

---

[65] Marco Civil da Internet: "Art. 2º A disciplina do uso da internet no Brasil tem como fundamento o respeito à liberdade de expressão, bem como: I - o reconhecimento da escala mundial da rede; II - os direitos humanos, o desenvolvimento da personalidade e o exercício da cidadania em meios digitais; III - a pluralidade e a diversidade; IV - a abertura e a colaboração; V - a livre iniciativa, a livre concorrência e a defesa do consumidor; e VI - a finalidade social da rede."

[66] Sobre o conceito de censura prévia, ver: SARLET, Ingo Wolfgang; MARINONI, Luiz Guilherme; MITIDIERO, Daniel. *Curso de direito constitucional*. 4ª ed. São Paulo: Saraiva, 2015, *ebook*, pp. 795-796.

[67] Com base nas lições de Jónatas E. M. Machado, em sua monografia *Liberdade de expressão – Dimensões constitucionais da esfera pública no sistema social*, Daniel Sarmento afirma que o conceito restrito de censura seria "a restrição prévia à liberdade de expressão realizada por autoridades administrativas, que resulta na vedação à veiculação de um determinado conteúdo". Nesse caso, a censura provocada por órgãos administrativos seria qualificada como inconstitucional *prima facie*. Contudo, sustenta o autor que alargado o conceito de censura para abarcar restrições *a posteriori* que decorram do Legislativo, Judiciário ou mesmo de privados é possível que existam justificativas que tornem as medidas constitucionais, apesar de pesar contra elas presunção de inconstitucionalidade (SARMENTO, Daniel. *Comentários ao art. 5.º, incisos IV, V e IX...*, pp. 556-557).

[68] RODRIGUES JUNIOR, Otávio Luiz. Artigo 5º, incisos IV ao IX. In Paulo Bonavides e Jorge Miranda e Walber de Moura Agra (Coordenadores Científicos). Francisco Bilac Pinto Filho e Otavio Luiz Rodrigues Junior (Coordenadores Editoriais). *Comentários à Constituição Federal de 1988*. Rio de Janeiro: Forense, 2009, p. 108: "(...) o legislador constituinte afastar: a) o controle prévio ou posterior da divulgação do resultado das atividades intelectuais em sentido lato; b) a necessidade de obtenção de licença para a divulgação dessas atividades."

[69] BARROSO, Luís Roberto. Liberdade de expressão versus direitos da personalidade – Colisão de direitos fundamentais e critérios de ponderação. *Temas de direito constitucional*, t. III, pp. 105-106.

[70] MENDES, Gilmar. *Direitos fundamentais e controle de constitucionalidade*..., p. 647: "É fácil ver, assim, que o texto constitucional não excluiu a possibilidade de que se introduzam limitações à liberdade de expressão e de comunicação, estabelecendo, expressamente, que o exercício dessas liberdades há de se fazer com observância do disposto na Constituição. Não poderia ser outra a orientação do constituinte, pois, do contrário, outros valores, igualmente relevantes, quedariam esvaziados diante de um direito avassalador, absoluto e insuscetível de restrição."

[71] TOFFOLI, Dias. Fake News: desinformação e liberdade de expressão. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, p. 36: "O Supremo Tribunal Federal tem construído uma jurisprudência consistente

essa liberdade "deve ser exercida em harmonia com os demais direitos e valores constitucionais", não podendo "respaldar a alimentação do ódio, da intolerância e da desinformação", que são situações representativas do "exercício abusivo desse direito". De acordo com o ministro, essa teria sido uma das razões pelas quais o STF "manteve condenação de um escritor e editor julgado pelo crime de racismo por publicar, vender e distribuir material antissemita", fazendo menção ao julgamento do caso Ellwanger (HC 82.424/STF). A respeito desse caso, o ministro acrescentou também que a garantia da liberdade de expressão "foi afastada em nome dos princípios da dignidade da pessoa humana e da igualdade jurídica".[72]

A liberdade de expressão garantida pelo artigo 5º da Constituição de 1988 está profundamente relacionada à cidadania (artigo 1º, II), à dignidade da pessoa humana (artigo 1º, III) e ao pluralismo político (artigo 1º, V), fundamentos da República.[73] A liberdade de expressão, então, além de obedecer aos limites legais contra práticas ilícitas, deve promover a justiça social, reduzir desigualdades e fortalecer a dignidade humana.[74]

---

em defesa da liberdade de expressão: declarou a inconstitucionalidade da antiga Lei de Imprensa, por possuir preceitos tendentes a restringir a liberdade de expressão de diversas formas (ADPF 130, DJe de 6/11/2009); afirmou a constitucionalidade das manifestações em prol da legalização da maconha, tendo em vista o direito de reunião e o direito à livre expressão de pensamento (ADPF 187, DJe de 29/5/14); dispensou diploma para o exercício da profissão de jornalismo, por forma da estreita vinculação entre essa atividade e o pleno exercício das liberdades de expressão e de informação (RE 511.961, DJe de 13/11/09); determinou, em ação de minha relatoria, que a classificação indicativa das diversões públicas e dos programas de rádio e TV, de competência da União, tenha natureza meramente indicativa, não pode ser confundida com licença prévia (ADI 2404, DJe de 1/8/17) - para citar apenas alguns casos."

[72] *Idem*.

[73] BARROSO, Luís Roberto; BARROSO, Luna van Brussel. Democracia, mídias sociais e liberdade de expressão: ódio, mentiras e a busca da verdade possível. *Direitos Fundamentais & Justiça*, Belo Horizonte, ano 17, n. 49, p. 285-311, jul./dez. 2023, p. 296: "A liberdade de expressão é um direito fundamental incorporado em praticamente todas as constituições contemporâneas e, em muitos países, é considerada uma liberdade preferencial, que deve prevalecer *prima facie* quando em confronto com outros valores. Várias razões procuram justificar a sua proteção especial, incluindo: (i) *a busca pela verdade possível* em uma sociedade aberta e plural; (ii) como *elemento essencial para a democracia*, pois permite a livre circulação de ideias, informações e pontos de vista que informam a opinião pública e o voto; e (iii) como *elemento essencial da dignidade humana*, permitindo a expressão da personalidade de cada pessoa."

[74] BINENBOJM, Gustavo. Meios de comunicação de massa, pluralismo e democracia deliberativa. As liberdades de expressão e de imprensa nos Estados Unidos e no Brasil. *Revista da EMERJ*, v. 6, n. 23, 2003, p. 373: "(...) o art. 5°, incisos IV e IX e o art. 220, *caput*, §§ 1° e 2° asseguram tais liberdades, com o banimento de qualquer censura política, ideológica e artística. A reconquista de tais garantias liberais merece ser celebrada e preservada. Ocorre que, de parte sua preocupação com a dimensão individual e *defensiva* da liberdade de expressão (entendida como proteção contra ingerências indevidas do Estado na livre formação do pensamento dos cidadãos), o constituinte atentou também para a sua dimensão transindividual e *protetiva*, que tem como foco o enriquecimento da qualidade e do grau de *inclusividade* do discurso público. É interessante notar que, ao contrário da Constituição dos Estados Unidos, a Constituição brasileira de 1988 contempla,

No direito contemporâneo, essa visão ganhou força com as Constituições intervencionistas do século XX, como a brasileira de 1988, que, além de proteger direitos individuais, estabeleceram metas sociais a serem perseguidas pelos cidadãos. Nos séculos XVIII e XIX, a ideia predominante era de uma liberdade de expressão que enfatizava a liberdade individual e afastava interferências do Estado nos assuntos privados. Hoje, a ordem constitucional destaca a liberdade aliada à solidariedade, exigindo que a autonomia privada atenda também a valores coletivos e ao interesse público, indo além de uma visão puramente individualista.[75]

Ingo Wolfgang Sarlet, Luiz Guilherme Marinoni e Daniel Mitidiero afirmam que o regime jurídico conduz à interpretação moderada do conceito de censura (deixando equivalê-la à restrição), com o objetivo de não tornar, por ricochete, a liberdade de expressão um direito absoluto, insensível às ponderações e restrições do próprio ordenamento jurídico em que se insere. Em suas palavras:

> O problema de uma definição demasiadamente ampla de censura, como abarcando toda e qualquer restrição à liberdade de expressão, é de que ela acabaria por transformar a liberdade de expressão em direito absoluto, o que não se revela como sustentável pelo prisma da equivalência substancial e formal entre a liberdade de expressão e outros bens fundamentais, pelo menos a dignidade da pessoa humana e os direitos de personalidade. Por outro lado, tomando-se também a liberdade de expressão como abarcando as diversas manifestações que lhe são próprias, a liberdade de manifestação do pensamento, a liberdade de comunicação e de informação (relacionadas com a liberdade de imprensa), a liberdade de expressão artística, apenas para citar as mais importantes, verifica-se que uma distinção entre censura e outras modalidades de restrição (que poderão, a

ela mesma, os princípios que devem ser utilizados no sopesamento das dimensões defensiva e protetiva da liberdade de expressão."

[75] FISS, Owen M. *The irony of free speech.* Cambridge: Harvard University Press, 1998, p. 18: "The concern is not simply with the social standing of the groups that might be injured by the speech whose regulation is contemplated. Rather, the concern is with the claims of those groups to a full and equal opportunity to participate in public debate—the claims of these groups to their right to free speech, as opposed to their right to equal protection. The state, moreover, is honoring those claims not because of their intrinsic value or to further their self-expressive interests but only as a way of furthering the democratic process. The state is trying to protect the interest of the audience—the citizenry at large—in hearing a full and open debate on issues of public importance". Tradução: "A preocupação não está simplesmente com o status social dos grupos que poderiam ser prejudicados pela fala cuja regulação está sendo contemplada. Em vez disso, a preocupação está com as reivindicações desses grupos por uma oportunidade plena e igual de participar do debate público — as reivindicações desses grupos pelo seu direito à liberdade de expressão, em oposição ao seu direito à proteção igualitária. O Estado, além disso, está honrando essas reivindicações não por seu valor intrínseco ou para promover seus interesses de autoexpressão, mas apenas como uma maneira de promover o processo democrático. O Estado está tentando proteger o interesse da audiência — a cidadania em geral — em ouvir um debate completo e aberto sobre questões de interesse público."

> depender do caso, ser constitucionalmente justificadas) é necessária até mesmo para preservar as peculiaridades de cada modalidade da liberdade de expressão.[76]

Disposto o regime jurídico da liberdade de expressão e sua possível restrição para resgatá-la de seu próprio exercício abusivo, será necessário avaliar a jurisprudência do Superior Tribunal de Justiça e os recentíssimos debates no âmbito do Supremo Tribunal Federal a respeito do exercício da liberdade de expressão na internet, considerando a regra de responsabilidade civil prevista no Marco Civil da Internet e o contexto de vácuo normativo sobre a moderação de conteúdo.

1.5. A regulação da liberdade de expressão nas redes sociais pela norma vigente: Marco Civil da Internet

O Marco Civil da Internet brasileiro ("MCI") tornou-se o regime jurídico do mundo digital ao estabelecer seus princípios, direitos e deveres para o uso da internet. Em seu artigo 19, o Marco Civil da Internet estabelece a responsabilidade civil dos provedores de internet, aí inclusas as plataformas digitais (MCI, artigo 5, VII),[77] por conteúdo gerado por terceiro com a finalidade expressa de assegurar a liberdade de expressão e impedir a censura:

> Artigo 19. 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.

O regime de responsabilização do MCI prevê que as plataformas digitais somente poderão ser responsabilizadas pela não retirada de conteúdo de terceiro após ordem judicial prévia e específica (*Judicial Notice and Takedown*), não bastando a notificação da vítima (*Notice and Takedown*).[78] Essa regra é

---

[76] SARLET, Ingo Wolfgang; MARINONI, Luiz Guilherme; MITIDIERO, Daniel. *Curso de direito constitucional*..., pp. 795-796.

[77] Marco Civil da Internet: "Art. 5º Para os efeitos desta Lei, considera-se: (...) VII - aplicações de internet: o conjunto de funcionalidades que podem ser acessadas por meio de um terminal conectado à internet (...)".

[78] O procedimento do *Notice and TakeDown* tem origem no Digital Millennium Copyright Act ("DMCA"), lei federal aprovada pelo Congresso dos Estados Unidos da América em 1998, que

excepcionada apenas nas hipóteses em que o conteúdo indicado consistir em imagens, vídeos e outros materiais que infrinjam direitos de autor ou contenham cenas de nudez ou de atos sexuais de caráter privado não autorizados, conforme dispõem o §2° do artigo 19[79] e o artigo 21 do MCI, respectivamente.[80]

Antes da vigência do MCI, a jurisprudência do Superior Tribunal de Justiça - STJ adotava o sistema *Notice and Takedown*, de modo que as plataformas poderiam ser consideradas corresponsáveis solidariamente com o causador do dano, caso não inviabilizasse o acesso ao conteúdo lesivo após a notificação da vítima.[81]

O anterior posicionamento do STJ, embora reconhecesse que o provedor não teria obrigação de exercer controle prévio do conteúdo postado por terceiro, tendia à imediata retirada do conteúdo apontado como ilegal, porque se orientou, na ausência de parâmetro legal, na fixação de um prazo de 24 (vinte e quatro) horas para retirada do conteúdo danoso após a notificação extrajudicial da vítima, sob pena de responsabilização.[82]

---

regula a tutela dos direitos autorais na internet. A *Section 512* do DMCA fornece proteção específica diante de alegações de violação de direitos autorais, trazendo o procedimento de *Notice and Takedown* para retirada imediata, pelos International Providers Services-IPS, de conteúdos violadores desses direitos. Por meio desse procedimento, a vítima poderia tomar providências em casos de violação de direitos autorais, criando-se um ambiente de autorregulação pelos usuários da internet. (Estados Unidos da America. *Digital Millennium Copyright Act*. Disponível em: <https://www.congress.gov/bill/105th-congress/house-bill/2281/text>. Acesso em 05 dez. 2024).

[79] Marco Civil da Internet: "Art. 19. (...) § 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."

[80] Marco Civil da Internet: "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."

[81] PEREIRA DE LIMA, Cíntia Rosa; FRANCO DE MORAES, Emanuele Pezati; PEROLI, Kelvin. *O necessário diálogo entre o Marco Civil da Internet e a Lei Geral de Proteção de Dados para a coerência do sistema de responsabilidade civil diante das Novas Tecnologias*..., p. 146.

[82] RECURSO ESPECIAL. CIVIL E PROCESSUAL CIVIL. RESPONSABILIDADE CIVIL. INTERNET. DANO MORAL. CRIAÇÃO DE PERFIS FALSOS E COMUNIDADES INJURIOSAS EM SÍTIO ELETRÔNICO MANTIDO POR PROVEDOR DE INTERNET. RELAÇÃO DE CONSUMO. AUSÊNCIA DE CENSURA. NOTIFICADO O PROVEDOR, TEM O PRAZO DE 24 HORAS PARA EXCLUIR O CONTEÚDO DIFAMADOR. DESRESPEITADO O PRAZO, O PROVEDOR RESPONDE PELOS DANOS ADVINDOS DE SUA OMISSÃO. PRECEDENTES ESPECÍFICOS DOS STJ. (...) 3. Polêmica em torno da responsabilidade civil por omissão do provedor de internet, que não responde objetivamente pela inserção no site, por terceiros, de dados ilícitos. 4. Impossibilidade de se impor ao provedor a obrigação de exercer um controle prévio acerca do conteúdo das informações postadas no site por seus usuários, pois constituiria uma modalidade de censura prévia, o que não é admissível em nosso sistema jurídico. 5. Ao tomar

Dessa forma, buscava-se estancar a disseminação do conteúdo indicado como ilícito pela vítima antes mesmo da confirmação da veracidade denúncia, que, posteriormente seria analisada, seja pelo provedor de internet com base em seus termos de uso, seja pelo próprio Poder Judiciário no momento da entrega da atividade jurisdicional final com a prolação da sentença de mérito.

A despeito de terem sido criados alguns parâmetros para a moderação de conteúdo por construção jurisprudencial, com a promulgação do MCI, verificou-se a alteração do entendimento do STJ que passou a observar a regra disposta no novo regime jurídico. É o que se extrai, inclusive, de trecho de julgado da lavra da ministra Nancy Andrighi sobre o tema:

> (...) 7. Com o advento da Lei 12.965/2014, o termo inicial da responsabilidade do provedor de aplicação foi postergado no tempo, iniciando-se tão somente após a notificação judicial do provedor de aplicação. 8. A regra a ser utilizada para a resolução de controvérsias deve levar em consideração o momento de ocorrência do ato lesivo ou, em outras palavras, quando foram publicados os conteúdos infringentes: (i) para fatos ocorridos antes da entrada em vigor do Marco Civil da Internet, deve ser obedecida a jurisprudência desta corte; (ii) após a entrada em vigor da Lei 12.965/2014, o termo inicial da responsabilidade da responsabilidade solidária do provedor de aplicação, por força do artigo 19 do Marco Civil da Internet, é o momento da notificação judicial que ordena a retirada de determinado conteúdo da internet. Recurso especial conhecido e provido.[83]

Além de alterar seu entendimento, o STJ acrescentou, com base no disposto no §1º do artigo 19 do Marco Civil da Internet – que prevê a necessidade de a ordem judicial ser "específica" –,[84] um requisito de validade da decisão judicial que determina a retirada de conteúdo *online*: o fornecimento do endereço eletrônico do conteúdo a ser retirado (*Uniform Resource Locator* – URL).[85]

---

conhecimento, porém, da existência de dados ilícitos em "site" por ele administrado, o provedor de internet tem o prazo de 24 horas para removê-los, sob pena de responder pelos danos causados por sua omissão. (...) (BRASIL. Superior Tribunal de Justiça. REsp n. 1.337.990/SP, Min. Rel. Paulo de Tarso Sanseverino, 3ª Turma, j. 21 ago. 2014).

[83] BRASIL. Superior Tribunal de Justiça. REsp n. 1.642.997/RJ, Minª. Relª. Nancy Andrighi, 3ª Turma, j. 12 set. 2017.

[84] De acordo com o §1º do artigo 19 do Marco Civil da Internet: "A ordem judicial de que trata o caput deverá conter, sob pena de nulidade, identificação clara e específica do conteúdo apontado como infringente, que permita a localização inequívoca do material."

[85] Nesse sentido: (i) BRASIL. Superior Tribunal de Justiça. REsp n. 1.568.935/RJ, Min. Rel. Ricardo Villas Bôas Cueva, Presidência do STJ, j. 5 abr. 2016; (ii) BRASIL. Superior Tribunal de Justiça. REsp n. 1.698.647/SP, Minª. Relª. Nancy Andrighi, 3ª Turma, j. 6 fev. 2018; e (iii) BRASIL. Superior Tribunal de Justiça. AgInt nos EDcl no REsp n. 1.402.112/SE, Min. Rel. Lázaro Guimarães, 4ª Turma, j. 19 jun. 2018.

Dessa forma, o sistema atual condiciona a responsabilidade civil (subjetiva) das plataformas digitais por omissão, como regra, à prévia e específica notificação judicial para retirada do conteúdo ilícito.[86] Essa regra somente será excepcionada nas hipóteses de "imagens, de vídeos ou de outros materiais contendo cenas de nudez ou de atos sexuais de caráter privado" não autorizados (MCI, artigo 21).

1.5.1. Vácuo normativo sobre a moderação de conteúdo

O sistema de regulação da internet estabelecido, até então, pelo Marco Civil da Internet não dispõe sobre parâmetros, critérios ou normas procedimentais sobre a moderação de conteúdo realizada pelas redes sociais.

Ao limitar o escrutínio público sobre a moderação de conteúdo *online* tão somente ao Poder Judiciário, característico por sua seletividade e limitação em expertise técnica, "a regra [estabelecida no MCI] opera como um regime de autorregulação, que deixa a cargo exclusivo da empresa privada uma vasta quantidade de decisões sobre parâmetros do exercício da liberdade de expressão *online*".[87] A esse respeito, adverte a doutrina:

> Na ausência de parâmetros legais, impera a definição de critérios pelas próprias plataformas, e por conseguinte, condicionados ao seu entendimento unilateral sobre o que cabe e o que não cabe na esfera pública digital.[88]

---

[86] MULHOLLAND, Caitlin. Responsabilidade civil indireta dos provedores de serviço de internet e sua regulação no Marco Civil da Internet. In José Renato Gaziero Cella, Aires Jose Rover, Valéria Ribas Do Nascimento (Coords.). *Direito e novas tecnologias*. Florianópolis: CONPEDI, 2015, pp. 479-502, p. 486: "Significa isso dizer que, no ordenamento brasileiro, o provedor de aplicação não tem o dever de verificar previamente e impedir o conteúdo a ser postado por terceiro (o que configuraria censura) porque ele não será responsabilizado posteriormente pelos danos causados pelo mesmo. Isto é, a responsabilidade pelo conteúdo gerado, postado e/ou disseminado na Internet recai primeiramente e, como regra, sobre aquele que diretamente realiza a conduta danosa, excluindo a responsabilidade do provedor em relação à vítima do dano. Mas, esta regra estabelece uma exceção bastante específica: o provedor será responsabilizado se, notificado judicialmente quanto ao conteúdo impróprio postado por terceiro, não providenciar a retirada do mesmo em prazo determinado. Verifica-se, assim, que para a responsabilização legal do provedor de aplicações em caso de conteúdo gerado por terceiros são requisitos necessários: 1) a existência de pedido de notificação judicial realizada por pessoa que alega a violação de seu direto (fundamental, autoral, intelectual, etc.); 2) a avaliação positiva, ainda que liminar e antecipada, pelo juiz quanto à potencial lesividade da conduta daquele que inseriu conteúdo; 3) a decisão liminar concedendo notificação pelo juiz ao provedor de aplicação indicando o conteúdo indevido a ser retirado e prazo para tal; 4) o descumprimento da decisão judicial da notificação para retirada."

[87] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais...*, p. 69.

[88] *Ibidem*, p. 70.

A norma de regência não traz mecanismos que acrescentem camadas de governança sobre a moderação de conteúdo, como transparência, devido processo ou critérios de desmonetização de conteúdo.[89]

O Marco Civil da Internet tampouco dispõe sobre o tratamento de conteúdos proporcional aos danos em potencial, porque não define parâmetros para uma curadoria de conteúdo voltada a população vulnerável, grupos minorizados ou a questões sensíveis, além de não assegurar a participação dos indivíduos e grupos afetados no debate sobre o escopo da liberdade de expressão.[90]

O sistema instituído pelo artigo 19 do MCI, embora possa ser considerado como uma garantia de maior segurança jurídica para evitar a remoção excessiva de conteúdo ao deixar as decisões sobre o escopo da liberdade da expressão ao Poder Judiciário,[91] é objeto de contundentes críticas de parte da doutrina civilista, que o considera um retrocesso no regime de responsabilização civil das plataformas digitais e na tutela dos direitos da personalidade no ambiente virtual.[92]

Por tratar especificamente do sistema de responsabilização das plataformas digitais por conteúdo de terceiros, sem estabelecer regras, parâmetros ou critérios para a moderação de conteúdo, o atual regime de regulação das plataformas digitais, previsto no Marco Civil da Internet, não é capaz, portanto, de garantir uma governança democrática de conteúdo nas redes sociais.

---

[89] *Ibidem*, p. 84.
[90] *Idem*.
[91] *Ibidem*, p. 68.
[92] De acordo com Anderson Schreiber, o retrocesso reside no fato de que, antes, a vítima só propunha ação judicial com o objetivo de responsabilizar o demandado como última medida, ao passo que, agora, obrigatoriamente, precisa propor a ação para "(...) pleitear a emissão de uma ordem judicial específica, para que, só então e apenas em caso de descumprimento da referida ordem judicial, a proprietária do *site* ou rede social possa ser considerada responsável" (SCHREIBER, Anderson. *Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro*..., p. 14). No mesmo sentido, Cíntia Rosa Pereira Lima afirma que a norma seria considerada um retrocesso, tanto por aumentar a judicialização dos conflitos quanto por gerar demora na exclusão do conteúdo da *internet* e, consequentemente, perpetuação do dano, "(...) na medida em que será acessado por milhares de pessoas e, em pouco tempo, o conteúdo se espalhará inviabilizando o retorno ao *status quo*". (PEREIRA DE LIMA, Cíntia Rosa; FRANCO DE MORAES, Emanuele Pezati; PEROLI, Kelvin. *O necessário diálogo entre o Marco Civil da Internet e a Lei Geral de Proteção de Dados para a coerência do sistema de responsabilidade civil diante das Novas Tecnologias*..., p. 146).

1.5.2. O julgamento da constitucionalidade do artigo 19 do Marco Civil da Internet pelo STF

A exigência de prévia e específica ordem judicial como marco inicial da responsabilização dos provedores deu ensejo à discussão da constitucionalidade do artigo 19 do MCI perante o STF no âmbito dos Recursos Extraordinários 1.037.396/SP (Tema 987)[93] e 1.057.258/MG (Tema 533)[94], sob a relatoria dos ministros Dias Toffoli e Luiz Fux, respectivamente.

Como os recursos estão sendo julgados conjuntamente, o ministro Dias Toffoli ("Toffoli") proferiu o voto inaugural para declarar a *inconstitucionalidade total* do artigo 19 do MCI e estabelecer uma nova abordagem acerca da regulação do discurso público nas redes sociais.[95] Em linhas gerais, o ministro Toffoli concluiu que o condicionamento da remoção de conteúdo à notificação judicial restringe indevidamente a responsabilização civil das plataformas digitais.

Ao proferir o voto inaugural, o ministro Toffoli propôs solução que envolvia a responsabilização dos provedores por danos causados por terceiros, nos casos em que não tomarem medidas adequadas em "prazo razoável" após notificação extrajudicial da vítima. Essa solução, aliás, aproxima-se daquela que era adotada pela jurisprudência do STJ antes da vigência do Marco Civil da Internet.

De forma mais extensiva, o ministro Toffoli estabeleceu a responsabilidade objetiva dos provedores quando recomendem, moderem ou impulsionem conteúdo que esteja previsto em um rol taxativo de crimes[96] ou nos casos de contas

---

[93] Tema 987/STF: "Discussão sobre a constitucionalidade do art. 19 da Lei n. 12.965/2014 (Marco Civil da Internet) que determina a necessidade de prévia e específica ordem judicial de exclusão de conteúdo para a responsabilização civil de provedor de internet, websites e gestores de aplicativos de redes sociais por danos decorrentes de atos ilícitos praticados por terceiros." (BRASIL. Supremo Tribunal Federal. Disponível em: <https://portal.stf.jus.br/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=5160549&numeroProcesso=1037396&classeProcesso=RE&numeroTema=987>. Acesso em 5 dez. 2024).

[94] Tema 533/STF: "Dever de empresa hospedeira de sítio na internet fiscalizar o conteúdo publicado e de retirá-lo do ar quando considerado ofensivo, sem intervenção do Judiciário." (BRASIL. Supremo Tribunal Federal. Disponível em: <https://portal.stf.jus.br/jurisprudenciarepercussao/verAndamentoProcesso.asp?incidente=5217273&numeroProcesso=1057258&classeProcesso=RE&numeroTema=533>. Acesso em 5 dez. 2024).

[95] BRASIL. Supremo Tribunal Federal. RE n. 1.057.258/MG, Min. Rel. Luiz Fux, Plenário, voto do Relator em 11 dez. 2024; e BRASIL. Supremo Tribunal Federal. RE n. 1.037.396/SP, Min. Rel. Dias Toffoli, Plenário, voto do Relator em 5 dez. 2024.

[96] O ministro Toffoli elencou o dever de retirar conteúdo que constituíssem *(i)* crimes contra o Estado Democrático de Direito, *(ii)* terrorismo, *(iii)* racismo, *(iv)* violência contra crianças, mulheres e grupos vulneráveis, *(v)* incitação ao suicídio, *(vi)* desinformação com potencial lesivo que levem à incitação à violência física, ameaça contra a vida ou atos de violências contra grupos vulneráveis ou que possam causar danos à integridade do processo eleitoral, *(vii)* infrações sanitárias, *(viii)* incitação à violência física ou sexual e *(ix)* tráfico de pessoas (*Idem*).

inautênticas, desidentificadas ou automatizadas.[97] E, ainda, estipulou a solidariedade das redes sociais com anunciantes ou patrocinadores no contexto de conteúdos publicitários.

A partir disso, o ministro Toffoli se baseou nos princípios da *boa-fé objetiva*, *função social* e *prevenção e mitigação de dano* para prever deveres anexos às redes sociais, como, por exemplo, procedimentos padronizados para a moderação de conteúdo.[98] Embora iniciada uma "regulação" (leia-se, precária) em torno da moderação de conteúdo da internet, modificando o regime jurídico do Marco Civil da Internet, o ministro fez um "apelo", em seu voto, aos Poderes Executivo e Legislativo, para que fosse estruturada uma regulação mais consistente a respeito da matéria.

Acompanhando o voto do ministro Toffoli, o ministro Luiz Fux fixou a tese do Tema 533/STF afastando a necessidade da prévia decisão judicial para que os provedores de internet sejam responsáveis pela "remoção imediata" e impôs um dever de monitoramento ativo, com a declaração de *inconstitucionalidade total* do artigo 19 do MCI. Confira-se:

> 1. A disposição do art. 19 do Marco Civil da Internet (Lei Federal nº 12.965/2014) não exclui a possibilidade de responsabilização civil de provedores de aplicações de internet por conteúdos gerados por terceiros nos casos em que, tendo ciência inequívoca do cometimento de atos ilícitos, seja porquanto evidente, seja

[97] Embora restrita às hipóteses de moderação e impulsionamento de conteúdos previstos em rol criminal específico e de contas inautênticas ou automatizadas, a estipulação de responsabilidade objetiva das plataformas digitais pressupõe a aplicação da teoria do risco, tal como aplicável aos veículos tradicionais de mídia, que, ao contrário das redes sociais, possuem controle editorial sobre o conteúdo publicado. No entanto, esse regime não é definido, até hoje, em nenhum país democrático, porque "a pretensão de que elas [as plataformas] fiscalizem tudo o que é postado em suas redes revela incompreensão essencial do ambiente digital e ameaça a própria sustentabilidade desse modelo de negócios". (BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., pp. 229-230).

[98] Em síntese, o ministro Toffoli defendeu a existência dos seguintes deveres: (a) manter atualizados e publicizar amplamente os termos de uso e regulamentos adicionais; (b) assegurar a autenticidade das contas e a identificação correta dos usuários, impedindo a criação ou bloqueando contas inautênticas, desidentificadas ou automatizadas; (c) elaborar códigos de conduta; (d) definir regras claras e procedimentos padronizados para a moderação de conteúdos, com ampla divulgação; (e) atualizar continuamente critérios e métodos de moderação, garantindo a publicidade dessas informações; (f) combater a desinformação e notícias fraudulentas, neutralizando redes artificiais de distribuição e identificando contas responsáveis para envio de dados às autoridades competentes; (g) monitorar riscos sistêmicos dos ambientes digitais, com produção e ampla divulgação de relatórios semestrais de transparência sobre riscos e medidas adotadas; (h) disponibilizar canais específicos, preferencialmente eletrônicos, para denúncias de conteúdos ilícitos ou ofensivos, com apuração prioritária; e (i) elaborar relatórios semestrais sobre a gestão e resolutividade de reclamações recebidas (BRASIL. Supremo Tribunal Federal. RE n. 1.057.258/MG, Min. Rel. Luiz Fux, Plenário, voto do Relator em 11 dez. 2024; e BRASIL. Supremo Tribunal Federal. RE n. 1.037.396/SP, Min. Rel. Dias Toffoli, Plenário, voto do Relator em 5 dez. 2024).

> porque devidamente informados por qualquer meio idôneo, não procederem à remoção imediata do conteúdo. 2. Considera-se evidentemente ilícito (item 1) o conteúdo gerado por terceiro que veicule discurso de ódio, racismo, pedofilia, incitação à violência, apologia à abolição violenta do Estado Democrático de Direito e apologia ao Golpe de Estado. Nestas hipóteses específicas, há para as empresas provedoras um dever de monitoramento ativo, com vistas à preservação eficiente do Estado.[99]

Abrindo divergência em relação aos votos dos ministros relatores, o ministro Luís Roberto Barroso ("Barroso") votou pela *inconstitucionalidade parcial* do artigo 19, aplicando a técnica da interpretação conforme à Constituição.[100] O ministro Barroso defendeu a ampliação da sistemática do artigo 21 do MCI, que prevê a retirada de conteúdos sexuais ou de nudez sem necessidade de ordem judicial, estendendo esse modelo para conteúdos específicos, como crimes em geral.[101]

O ministro Barroso propôs dois regimes de responsabilização: no primeiro, a notificação extrajudicial seria suficiente para que a plataforma remova conteúdos ilegais, sob pena de responsabilização, exceto para crimes contra a honra, que permanecem sob a sistemática do artigo 19. No segundo modelo, baseado no "dever de cuidado" (*duty of care*), as plataformas deveriam atuar proativamente para mitigar riscos sistêmicos, especialmente em casos graves, como pornografia infantil, terrorismo e abolição violenta do Estado Democrático de Direito.

---

[99] BRASIL. Supremo Tribunal Federal. RE n. 1.057.258/MG, Min. Rel. Luiz Fux, Plenário, voto do Relator em 11 dez. 2024; e BRASIL. Supremo Tribunal Federal. RE n. 1.037.396/SP, Min. Rel. Dias Toffoli, Plenário, voto do Relator em 5 dez. 2024.

[100] O voto de Barroso, proferido em 18 de dezembro de 2024, ainda não foi disponibilizado ao público por órgãos oficiais. Nesse sentido, remeta-se a CARVALHO, Luísa. *Ponto a ponto: entenda o voto de Barroso sobre o artigo 19 do Marco Civil da Internet*. JOTA, publicado em 20 dez. 2024. Disponível em: https://www.jota.info/stf/do-supremo/ponto-a-ponto-entenda-o-voto-de-barroso-sobre-o-artigo-19-do-marco-civil-da-internet. Acesso em 7 mar. 2025.

[101] MULHOLLAND, Caitlin. *Responsabilidade civil indireta dos provedores de serviço de internet e sua regulação no Marco Civil da Internet*..., p. 498: "O artigo 21 do Marco Civil estabelece um rito diferente do artigo 19 no que se refere ao procedimento de notificação para a retirada de material de conteúdo infringente gerado por terceiro. Este artigo traz hipótese específica de tutela da intimidade das pessoas retratadas em imagens, vídeos ou outros materiais contendo cenas de nudez ou atos sexuais de caráter privado que, sem sua autorização, têm exibido esse conteúdo na rede. Por se tratar de conteúdo com grave potencial danoso, entendeu o legislador que a sua propagação não autorizada dispensa a necessidade de notificação judicial para que o provedor seja obrigado a retirar o conteúdo da rede. Para que o material seja indisponibilizado basta notificação da pessoa retratada, considerando que esta notificação, portanto, poderá ser extrajudicial e, portanto, o procedimento, mais célere. Ademais, a notificação deve 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' (Artigo 21, parágrafo único)."

O ministro Barroso destacou que as plataformas precisam prevenir e reduzir danos sistêmicos associados a seus modelos de negócios, incluindo riscos à segurança e à estabilidade democrática. A atuação proativa inclui a remoção de conteúdos "gravemente nocivos" e a adoção de medidas concretas para minimizar violações de direitos. Além disso, o ministro defendeu a publicação de relatórios anuais de transparência pelas plataformas com mais de dez milhões de usuários, alinhados ao Regulamento de Serviços Digitais Europeu (DSA – *Digital Services Act*), detalhando dados sobre conteúdos ilícitos removidos, tempo de resposta, e erros sistêmicos, que podem embasar ações de responsabilização, inclusive por dano moral coletivo. Para anúncios e conteúdos impulsionados, o ministro determinou a responsabilidade direta das plataformas digitais, independentemente de notificação.

O voto também reiterou a necessidade de equilíbrio entre a liberdade de expressão e a responsabilidade das plataformas, rejeitando a responsabilidade objetiva sugerida por outros ministros. O ministro Barroso propôs a criação de um órgão regulador independente e multissetorial para fiscalizar as plataformas e assegurar a transparência. Em seu voto, fez expressa distinção entre tipos de provedores de internet, excluindo *marketplaces* das mudanças propostas, e reforçou que cabe ao Congresso Nacional avançar na regulação do setor, especialmente por meio do Projeto de Lei 2630/2020 ("PL das Fake News").[102] Assim, o ministro buscou apresentar um modelo para equilibrar os direitos fundamentais no ambiente digital, fortalecendo a proteção contra abusos e promovendo a estabilidade democrática.

Com o pedido de vista do ministro André Mendonça, o julgamento do STF ainda se encontra em aberto, não tendo sido retomado até a data do depósito da presente dissertação.

Os entendimentos perfilados pelos ministros da Corte Suprema não afastam o debate sobre a regulação das redes sociais no contexto brasileiro. Ao revés, ao preverem a responsabilização civil por conteúdo postado por terceiros, em outras

[102] Em maio de 2024, o Presidente da Câmara dos Deputados Arthur Lira declarou que o PL das Fake News não possuía mais condições de alcançar o consenso necessário para aprovação da lei, o que levou ao descarte do referido projeto de lei. Ver, para mais detalhes, BONIN, Robson. Lira diz que PL das Fake News está morto e anuncia grupo por novo texto. *Veja*, publicado em 09 mai. 2024. Disponível em: <https://veja.abril.com.br/coluna/radar/lira-diz-que-pl-das-fake-news-esta-morto-e-anuncia-grupo-por-novo-texto>. Acesso em 27 dez. 2024.

hipóteses que não aquela disposta no Marco Civil da Internet, esboçando novos deveres a serem cumpridos pelas plataformas digitais, os votos proferidos pelos ministros reforçam o dever de moderar das plataformas digitais e, com ele, a necessidade da implementação de uma regulação que contribua para garantir maior confiabilidade e transparência à moderação de conteúdo nas redes sociais.

1.6. Por que regular as redes sociais no Brasil?

Delineados a evolução legislativa da liberdade de comunicação e os entendimentos doutrinário e jurisprudencial sobre a liberdade de expressão no Brasil à luz do Marco Civil da Internet, denota-se que a regulação da moderação de conteúdo feita pelas plataformas digitais não conflita, *prima facie*, com o princípio da liberdade de expressão, previsto no artigo 5º, parágrafo IV, da Constituição Federal de 1988,[103] tampouco configura qualquer espécie de censura prévia.

Mas antes de analisar *como* regular, é preciso responder ao seguinte questionamento: por que regular as redes sociais?[104] Entender por que regular é fundamental para compreender os objetivos da regulação das redes sociais no Brasil, evitando, assim, a adoção de medidas sem pertinência jurídica em relação aos bens jurídicos que se pretende tutelar.

O objetivo central da regulação das redes sociais deve consistir, essencialmente, na criação de incentivos adequados para que as empresas privadas que controlam as plataformas digitais se tornem instituições confiáveis e que contribuam para promover uma esfera pública digital saudável,[105] congruente com os direitos fundamentais e valores dispostos na ordem jurídica brasileira.

---

[103] BRASIL. *Constituição da República Federativa de 1988*. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em 08 mar. 2025: "Art. 5º Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes: (...) IV - é livre a manifestação do pensamento, sendo vedado o anonimato (...)".

[104] BALKIN, Jack M. *How to Regulate (and Not Regulate) Social Media*…, p. 72.

[105] *Idem*.

1.6.1. Assegurar os valores da liberdade de expressão e promover uma esfera pública digital saudável

Conforme desenvolvido acima, as empresas privadas que controlam as redes sociais são instituições fundamentais para manter a integridade da esfera pública digital. A partir dessa noção, vislumbra-se a existência de uma função pública das redes sociais, que está diretamente relacionada à manutenção de uma esfera pública digital saudável.

Jack M. Balkin define três funções centrais a serem desempenhadas pelas redes sociais, para que seja mantida uma esfera pública saudável: *(i)* facilitar a participação pública em arte, política e cultura; *(ii)* organizar o discurso público para facilitar a comunicação; e *(iii)* realizar a curadoria da opinião pública, através de resultados de pesquisa individualizados e dos mecanismos de moderação estabelecidos nos seus termos de uso.[106_107]

O cumprimento da função pública pelas redes sociais possibilita a concretização de, ao menos, três valores da liberdade de expressão: *(i)* a *democracia política*, que permite a participação democrática na formação da opinião pública, ajudando a garantir que o Poder Público seja sensível à sua evolução e que as pessoas estejam informadas a respeito das questões de interesse público; *(ii)* a *democracia cultural*, que possibilita que indivíduos e grupos possam participar livremente do intercâmbio cultural sem discriminação; e *(iii) o crescimento e a disseminação do conhecimento*.[108]

Em outras palavras, a esfera pública digital é saudável e funciona regularmente quando auxilia os indivíduos e grupos a realizar esses três valores

---

[106] O autor alerta que a curadoria feita pelas redes sociais abrange não apenas a retirada e reorganização de conteúdo, mas também a regulação da velocidade de propagação e alcance do conteúdo (*Ibidem*, p. 75).

[107] *Idem*: "I argue that social media have three central functions: First, social media facilitate public participation in art, politics, and culture. Second, social media organize public conversation so people can easily find and communicate with each other. Third, social media curate public opinion, not only through individualized search results and feeds, but also through enforcing community standards and terms of service. Social media curate not only by taking down or rearranging content, but also by regulating the speed of propagation and the reach of content". Tradução: "Defendo que as mídias sociais têm três funções centrais: Primeiro, as mídias sociais facilitam a participação pública na arte, na política e na cultura. Em segundo lugar, as mídias sociais organizam as conversas públicas para que as pessoas possam encontrar e se comunicar facilmente umas com as outras. Em terceiro lugar, a mídia social faz a curadoria da opinião pública, não apenas por meio de resultados de pesquisa e *feeds* individualizados, mas também por meio da aplicação de padrões comunitários e termos de serviço. A mídia social faz a curadoria não apenas retirando ou reorganizando o conteúdo, mas também regulando a velocidade de propagação e o alcance do conteúdo".

[108] *Ibidem*, p. 77.

centrais da liberdade de expressão. E, em sentido contrário, a esfera pública digital disfuncional causa prejuízos à democracia política e cultural e à disseminação do conhecimento. Uma esfera pública saudável demanda, portanto, mais do que o cumprimento da vedação à censura prévia.[109]

A realização dos valores da liberdade de expressão pressupõe a inovação, curadoria e disseminação do conhecimento por instituições intermediárias que sejam profissionais e em que o público confie. Na ausência dessas instituições, a liberdade de expressão se torna uma guerra retórica de todos contra todos, que prejudica as democracias política e cultural, bem como a disseminação do conhecimento que a liberdade de expressão deve promover.[110]

Para isso, as redes sociais devem se tornar instituições confiáveis e responsáveis e cumprir normas profissionais e públicas que orientem a forma como produzem, organizaram e distribuem conhecimento e opiniões para construir uma esfera pública saudável e vibrante.[111]

Embora advogados, pesquisadores e funcionários das redes sociais tenham considerável experiência sobre os sistemas de moderação de conteúdo e como eles funcionam, muitas vezes as plataformas não aplicam adequadamente seus próprios termos de uso. No contexto da moderação de conteúdo, pesquisadores apontam que, além dos problemas da não remoção de conteúdo ilícitos, as redes sociais cometem "remoção excessiva" para evitar os riscos de sua posterior responsabilização.[112_113]

Levando em consideração essa noção empírica, Daphne Keller e Paddy Leerssen estabelecem três objetivos concorrentes para a regulação das redes sociais: *(i)* evitar a ocorrência de danos aos usuários; *(ii)* proteger discursos e informações legais e válidos; e *(iii)* assegurar a inovação.[114] Assim, explicam os autores que o

---

[109] *Idem*.
[110] *Ibidem*, p. 79.
[111] *Idem*.
[112] Daphne Keller e Paddy Leerssen destacam, a título exemplificativo, que as redes sociais têm removido informações que vão desde reportagens e vídeos documentando brutalidade policial no Equador até a cobertura da mídia sobre investigações de fraude nos Estados Unidos, críticas a organizações religiosas e reportagens científicas (KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* …, p. 4.)
[113] *Idem*.
[114] *Ibidem*, pp. 7-8: "At a high level, Intermediary Liability laws must balance three, often competing goals. The legal details in national law typically reflect lawmakers' judgment about how best to balance them. One goal is to *prevent harm*. Generally, the better job a law does of incentivizing platforms to take down illegal or otherwise harmful content, the more it will serve this goal. Another, often competing goal is to *protect lawful online speech and information.* A law that requires aggressive policing by platforms may run afoul of this goal, leading platforms to take down lawful

equilíbrio para atingimento os referidos objetivos, por meio da regulação, dependem dos valores nacionais e das escolhas políticas:

> O equilíbrio de prioridades entre esses três objetivos é uma questão de valores nacionais e escolhas políticas. Mas a questão de quais regras legais específicas servirão, na prática, a cada objetivo é, em parte, empírica, ligada às práticas do mundo real das plataformas que respondem às exigências e aos incentivos da lei.[115]

De toda forma, uma legislação que estabeleça incentivos para a remoção de conteúdo ilegal provavelmente atingirá esses objetivos. Em sentido oposto, leis que exijam um policiamento agressivo podem levar as plataformas a removerem excessivamente discursos legais para evitar riscos legais e, ainda, causar excessivos dispêndios à moderação de conteúdo, com possíveis impactos sobre a livre concorrência entre as plataformas já consolidadas e as emergentes.

1.6.1.1. Regulação estrangeira: diferentes modelos de tutela da liberdade de expressão

A maioria das legislações nacionais prevê, de forma intencional, imunidade de responsabilidade civil às plataformas em caso de conteúdos danosos postados por terceiros, desde que não contribua na sua criação.

O *Communications Decency Act* ("CDA" – Lei de Decência das Comunicações), lei federal norte-americana, fornece proteção legal aos provedores

---

and valuable speech in order to avoid legal risk. A third goal is to *promote innovation*. Early Intermediary Liability laws were conceived in part as means to protect nascent industries. Today, Intermediary Liability laws may profoundly affect competition between incumbent platforms and startups." Tradução: "Em um nível elevado, as leis de responsabilidade civil do intermediário devem equilibrar três objetivos, muitas vezes concorrentes. As escolhas legais na legislação nacional normalmente refletem o julgamento dos legisladores sobre a melhor forma de equilibrá-los. Um objetivo é evitar danos. Em geral, quanto melhor for o trabalho de uma lei para incentivar as plataformas a remover conteúdo ilegal ou prejudicial, mais ela atenderá a esse objetivo. Outro objetivo, muitas vezes concorrente, é proteger o discurso e as informações *online* legais. Uma lei que exija um policiamento agressivo por parte das plataformas pode entrar em conflito com esse objetivo, levando as plataformas a retirarem do ar discursos legais e valiosos para evitar riscos legais. Um terceiro objetivo é promover a inovação. As primeiras leis de Responsabilidade Intermediária foram concebidas, em parte, como meios de proteger os setores nascentes. Atualmente, as leis de Responsabilidade do Intermediário podem afetar profundamente a concorrência entre as plataformas estabelecidas e as startups."

[115] *Ibidem*, p. 7: "The balance of priorities between these three goals is a matter of national values and policy choices. But the question of what specific legal rules will, in practice, serve each goal is in part an empirical one, tied to the real-world practices of platforms responding to the law's requirements and incentives".

online (IPS – *International Providers Services*) por conteúdo de terceiros (*Section 230*).[116]-[117] O CDA garante autonomia e independência às plataformas para decidir qual discurso deve ou não ser moderado, com propósito específico de incentivá-las a aperfeiçoar a moderação de conteúdo, por meio da autorregulação.[118] Em outros termos, a própria legitimidade das plataformas digitais para a moderação de conteúdo decorre da imunidade prevista na *Section* 230 do CDA.[119]

A *Section* 230 do CDA tem sido utilizada em casos como, por exemplo: *(a)* quando alguém publica um *tweet* difamatório, a rede social (*v.g.* Twitter) não poderia ser responsabilizada pelo conteúdo da postagem; *(b)* quando um usuário do *YouTube* publica um vídeo que viola direitos autorais, o *YouTube* poderia remover o vídeo sem ser responsabilizado legalmente; e *(c)* quando uma plataforma de mídia

---

[116] A *Section* 230 do CDA dispõe que: *(a)* os provedores de serviços *online* e plataformas de *internet* não são considerados editores ou publicadores dos conteúdos publicados por seus usuários; e *(b)* os provedores de serviços *online* e plataformas de *internet* podem remover ou restringir o acesso a conteúdo que considerem obscenos, lascivos, violentos ou de outra forma objetiváveis, desde que de boa-fé, sem que isso implique em responsabilidade, estabelecendo-se assim uma imunidade dos provedores nesses casos. (ESTADOS UNIDOS DA AMÉRICA. *Communications Decency Act*. Disponível em: <https://en.wikisource.org/wiki/Telecommunications_Act_of_1996#TITLE_V%E2%80%94OBSCENITY_AND_VIOLENCE >. Acesso em 22 dez. 2024).

[117] O contexto jurisprudencial anterior à promulgação do CDA sugeria que as plataformas poderiam ser responsabilizadas por conteúdo difamatório se realizassem qualquer controle editorial sobre o discurso postado por terceiro, conforme se extrai da análise dos *leading cases*: (i) *Cubby Inc. v. CompuServ*; e (ii) *Stratton Oakmont v. Prodigy Services Co*. No primeiro caso, houve postagem difamatória em jornal eletrônico armazenado em servidor do *CompuServ*, provedor de *internet* à época. Na Corte de Nova Iorque, o pleito foi rejeitado, por entender que o *CompuServ* deveria ser equiparado a um mero distribuidor da informação difamatória, isto é, tal como seria uma banca de jornal, livraria, biblioteca. Consequentemente, o provedor não poderia ser considerado o editor da informação, uma vez que não controlava o que era produzido (*Cubby, Inc. v. CompuServe Inc*., 776 *F. Supp*. 135. United States District Court, S.D. New York. Oct. 29, 1991). No segundo caso, houve postagem de mensagens difamatórias em *bulletin board* (BBS) da rede *Prodigy*. Diferentemente do primeiro caso, a decisão foi no sentido de admitir o pleito da causa, porquanto tenha se entendido que a rede possuía controle editorial, considerando que o provedor afirmava que a rede era "*Family oriented*", utilizando-se de meios para monitorar e controlar as informações. Esse posicionamento fez com que o tribunal entendesse que o provedor deveria ser equiparado ao editor primário da informação difamatória. Assim, firmou-se a tese de que o provedor pode ser responsabilizado quando assume, expressamente, um dever de controle ou de fiscalização do conteúdo, e desde que exista a possibilidade técnica de exercer efetivamente esse controle (*Stratton Oakmont, Inc. v. Prodigy Services Co*. 1995 WL 323710 at *5 (N.Y. Supr. Ct. May 23, 1995). Kate Klonick explica que esses casos "criaram um forte desestímulo para que os intermediários online expandissem seus negócios ou moderassem o conteúdo ofensivo e ameaçaram o cenário em desenvolvimento da Internet", tendo sido o CDA promulgado justamente, nesse contexto, para incentivar e proteger a remoção de material ofensivo pelas plataformas por meio da imunidade nele prevista (KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.605).

[118] *Idem*.

[119] *Ibidem*, p. 1.602.

social, como o *Facebook*, remove uma postagem que viola seus termos de serviço, ela não poderia ser responsabilizada pelo conteúdo da postagem.[120]

Esse modelo também foi adotado pelo legislador brasileiro ao estabelecer o regime previsto no Marco Civil da Internet. Conforme explicado acima no subcapítulo 1.5, a lei brasileira estabelece ampla imunidade às plataformas digitais, que somente serão responsabilizadas por conteúdo de terceiros se houver prévia decisão judicial que determine a retirada de conteúdo ilegal, não sendo suficiente, para tanto, o conhecimento prévio a respeito do conteúdo ilegal.

Outros modelos regulatórios preveem a possibilidade de responsabilização quando as plataformas tomam conhecimento do conteúdo ilegal. Em boa parte desses países, as redes sociais devem excluir qualquer conteúdo difamatório ou que contenha propaganda terrorista ao tomar conhecimento de sua existência, sob pena de responsabilização. Existem, nesse contexto, divergências sobre o que seria o conhecimento exigido para que as plataformas ajam proativamente.[121]

Existem ainda modelos de regulação que disciplinam procedimentos a serem aplicados no contexto da moderação de conteúdo. O *Digital Millennium Copyright Act* (DMCA – Lei dos Direitos Autorais do Milênio Digital), lei estadunidense que regula a tutela de direitos autorais, estabelece pré-requisitos formais para "notificações" de detentores de direitos (*Notice and Takedown*),[122]

---

[120] A partir dessas premissas, foi julgado o *leading case Kenneth M. Zenan v. American Online Incorporated – AOL* (*Zeran v. America Online*, Inc., 985 F. Supp. 1124 (E.D. Va. 1997)). Nesse caso, Kenneth moveu ação judicial contra *America Online Incorporated*, em virtude de um "trote" realizado por um usuário anônimo que se passou por Kenneth e veiculava mensagens de cunho difamatório ao anunciar camisetas comemorativas de atentado terrorista de Oklahoma City. Na época, embora a *American Online Incorporated* tenha sido avisada da natureza falsa e prejudicial do trote, não apagou as mensagens nem restringiu as postagens. E, a despeito disso, a demanda foi rejeitada sumariamente pela Corte Distrital da Virgínia, por entender que o provedor não pode ser demandado por qualquer tipo de ação ou ordem judicial, ainda que não removam o conteúdo difamatório. Firmou-se, assim, a tese de que a não retirada de postagem após a notificação de usuário da internet alvo de conteúdo difamatório não implicaria em qualquer responsabilidade – direta ou indireta – da plataforma digital, na forma do disposto na *Section 230* do CDA.

[121] *Ibidem*, p. 8.

[122] Ao estabelecer um procedimento de *Notice and Takedown* para endereçar alegações de violação de direitos autorais na *internet*, o DMCA disciplinou a matéria sob os seguintes aspectos: *(a)* em regra, o provedor de *internet* somente pode ser responsabilizado quando tem o conhecimento real da infração (*actual knowledge*); *(b)* o provedor de *internet* passa a ter um conhecimento presumido de sua existência (*constructive knowledge*); e *(c)* os provedores de *internet* devem ser previamente notificados a fim de possibilitar eventual responsabilização (BINICHESKI, Paulo Roberto. *Responsabilidade Civil dos Provedores de Internet*. Curitiba: Juruá, 2011, p. 80).

etapas de "contranotificação" por usuários acusados e, ainda, penalidades para notificações de má-fé contra discursos legais.[123]-[124]

De forma semelhante, os Princípios de Manila,[125]–[126] um conjunto de regras padrão de responsabilidade das plataformas endossado por grupos internacionais da sociedade civil, apoiado pela literatura de direitos humanos – e que influenciou a previsão do MCI de responsabilidade condicionada à prévia ordem judicial –, elenca proteções procedimentais que incluem requisitos de transparência pública

[123] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*…, p. 8.

[124] O DMCA estabelece, na *Section* 512(c), (1), (2) e (3), o procedimento de *Notice and TakeDown* nos seguintes termos: "Notice — Rightsholder sends notice to online service provider regarding infringing material that appears on the online service provider's system. Remove Access to Material — Online service provider must act expeditiously to remove or disable access to the infringing material. Notify User — Online service provider must then promptly notify the user that originally uploaded the material that it has been removed. Counter-notice — User may submit a counter-notice requesting the reinstatement of the material, if the user believes the removal was due to a mistake or misidentification. Restore Access or Initiate Court Action — Online service provider must restore access to the material after no less than 10 and no more than 14 business days, unless the original notice sender informs the service provider that it has filed a court action against the user". Tradução: "Aviso — O detentor dos direitos envia um aviso ao provedor de serviços *online* sobre material infringente que aparece no sistema do provedor de serviços *online*. Remoção de Acesso ao Material — O provedor de serviços *online* deve agir prontamente para remover ou desabilitar o acesso ao material infringente. Notificação ao Usuário — O provedor de serviços *online* deve então notificar prontamente o usuário que originalmente enviou o material de que foi removido. Contranotificação — O usuário pode enviar uma contranotificação solicitando a reinstauração do material, caso acredite que a remoção ocorreu por engano ou má identificação. Restaurar o Acesso ou Iniciar Ação Judicial — O provedor de serviços *online* deve restaurar o acesso ao material após no mínimo 10 e no máximo 14 dias úteis, a menos que o remetente do aviso original informe ao provedor de serviços que moveu uma ação judicial contra o usuário." (ESTADOS UNIDOS DA AMÉRICA. *Digital Millennium Copyright Act*. Disponível em: <Text - H.R.2281 - 105th Congress (1997-1998): Digital Millennium Copyright Act | Congress.gov | Library of Congress>. Acesso em 22 dez. 2024).

[125] Os Princípios de Manila reúnem seis orientações/recomendações a respeito da responsabilização das plataformas digitais, com o expresso "objetivo de proteger a liberdade de expressão e criar um ambiente propício à inovação, que equilibre as necessidades dos governos e de outras partes interessadas, grupos da sociedade civil de todo o mundo se reuniram para propor essa estrutura de salvaguardas básicas e práticas recomendadas", que se "baseiam em instrumentos internacionais de direitos humanos e em outras estruturas jurídicas internacionais". (ACESS NOW *et al*. *Princípios de Manila sobre Responsabilidade de Provedores*. Disponível em: <https://manilaprinciples.org/index.html>. Acesso em 15 dez. 2024).

[126] O modelo de responsabilidade recomendado nos Princípios de Manila envolve a responsabilização após a notificação judicial para remoção de conteúdo: "[o]s intermediários não devem ser obrigados a restringir o conteúdo, a menos que uma ordem tenha sido emitida por uma autoridade judicial independente e imparcial que tenha determinado que o material em questão é ilegal.". Original: "Intermediaries must not be required to restrict content unless an order has been issued by an independent and impartial judicial authority that has determined that the material at issue is unlawful." (*Idem*). Conforme já mencionado, esse sistema foi adotado pelo Brasil após a promulgação do Marco Civil da Internet e é adotado nos Estados Unidos da América (DCA), exceto para casos de violação a direitos autorais (DMCA).

para esclarecer erros, vieses ou abusos nos sistemas de notificação e retirada de conteúdo.[127]

A regulação alemã (NetzDG – *Netzwerkdurchsetzungsgesetz*), além de regras gerais de responsabilidade civil e de obrigações às redes sociais de promover a retirada em curto prazo de conteúdos ilícitos, conforme crimes tipificados no Código Penal alemão (StGB – Strafgesetzbuch), estabelece de forma prioritária a transparência sobre as denúncias recebidas e remoções de conteúdo.[128]-[129]

Martin Eifert, atual ministro do Tribunal Constitucional Federal alemão, considera que a lei alemã exige: "um 'procedimento efetivo e transparente' para o trato com as reclamações sobre conteúdos ilícitos que principalmente deva garantir o cumprimento à conformidade com prazos para o apagamento ou retirada de conteúdos ilícitos", estrutura normativa de obrigações por ele intitulada de *compliance system* (sistema de conformidade).[130]

No Reino Unido, discute-se atualmente a Lei de Segurança Online (*Online Safety Bill*), que estabelece às plataformas digitais uma categoria específica de obrigação conhecida como "dever de cuidado" (*duty of care*), que é voltada aos processos que compõem os modelos de negócios das plataformas digitais a fim de torná-los mais seguros.[131] Esse dever de cuidado também está previsto em uma das principais propostas legislativas de regulação no Brasil, o PL das Fake News (Projeto de Lei 2.630/2020),[132] além de ter sido mencionado no recente voto do ministro Barroso no julgamento da constitucionalidade do artigo 19 do MCI.

---

[127] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*…, p. 8.

[128] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 76.

[129] Como parte fundamental da solução de combate à propagação em massa de notícias falsas, de desinformação e do discurso de ódio no meio virtual, o diploma legal alemão estabeleceu, na Seção 2, a obrigação de que as redes sociais publiquem, semestralmente, relatórios para informar como "lidaram com as denúncias dos usuários acerca de conteúdos ilegais", sendo tal exigência "uma das mais elogiadas da lei", por garantir "maior transparência" e contribuir para o fornecimento de "material para estudo da temática". (BREGA, Gabriel Ribeiro. *A regulação de conteúdo nas redes sociais: uma breve análise comparativa entre o NetzDG e a solução brasileira*. Revista GV, São Paulo, v.19, e2305, 2023, p. 14. Disponível em: <https://doi.org/10.1590/2317-6172202305>. Acesso em 22 dez. 2024).

[130] EIFERT, Martin. A Lei Alemã para a melhoria da aplicação da Lei nas Redes Sociais (NetzDG) e regulação da plataforma. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.*). Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, p. 189.

[131] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 76.

[132] No artigo 6º, inciso II, do Projeto de Lei 2.630/2020, que visa instituir a "Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet", estipulou-se a responsabilidade civil

O Regulamento de Serviços Digitais Europeu (DSA – *Digital Services Act*), aprovado pelo Parlamento Europeu no final de 2022, possui, por sua vez, uma agenda mais abrangente, que incluem medidas concorrenciais e de proteção do direito dos consumidores.[133] Quanto à curadoria de conteúdo, o DSA manteve o anterior regime de responsabilidade da Diretiva de Comércio Eletrônico (*EU eCommerce Directive* ou Diretiva n. 2000/31/CE), baseado na responsabilização condicionada ao conhecimento sobre o conteúdo ilícito (*Notice and Takedown*), mas introduziu "camadas de governança", que englobam obrigações de transparência,[134] garantias de contraditório prévio à remoção de conteúdo e a análise de risco sistêmico.[135]

Não se ignora que a regulação pode perpassar caminhos que procurem fomentar o surgimento de um número maior de redes sociais por meio de medidas antitruste e concorrenciais que incentivem a inovação e de leis de proteção à privacidade e ao consumidor[136] ou, ainda, iniciativas que visam a coibir a recomendação algorítmica e o impulsionamento de conteúdos ilícitos, através da desmonetização de conteúdo.[137]

---

solidária das plataformas digitais em caso de descumprimento das obrigações de dever de cuidado previstas no artigo 11, na Seção III, do referido projeto legislativo. O artigo 11 do PL das Fake News prevê um dever de atuação diligente dos provedores de internet "para prevenir e mitigar práticas ilícitas no âmbito de seus serviços, envidando esforços para aprimorar o combate à disseminação de conteúdos ilegais gerados por terceiros, que possam configurar" determinados tipos penais previstos na legislação brasileira ali listados. (SILVA, Orlando. *Relatório final do Relator*, Deputado Orlando Silva. Apresentado em 31 de março de 2022, no Plenário da Câmara dos Deputados. Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=PRLP+1+%3D%3E+PL+2630/2020>. Acesso em 15 dez. 2024).

[133] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 77: O DSA é composto por um regulamento direcionado a serviços, que abrange disposições que refletem diversas tendências regulatórias mencionadas aqui, além de um Regulamento de Mercados Digitais (DMA), focado em questões de concorrência e direitos dos consumidores. Os objetivos desse regulamento estão voltados ao combate da concentração de poder das plataformas digitais em suas várias manifestações.

[134] Daphne Keller alerta para o risco de "remoção excessiva" criado pelo DSA na União Europeia, citando no mesmo contexto o DMCA nos EUA, porque essas regulações exporiam "as plataformas a litígios dispendioso e possível responsabilidade pelo discurso do usuário, que elas podem evitar simplesmente retirando mais discurso – mesmo que o discurso seja legal". (KELLER, Daphne. *Platform Transparency and the First Amendment* (March 3, 2023), p. 12. Disponível em: <https://ssrn.com/abstract=4377578>. Acesso em 3 jan. 2025).

[135] *Idem*.

[136] BALKIN, Jack M. *How to Regulate (and Not Regulate) Social Media*…, p. 72.

[137] Além de alertar para o problema da recomendação e do impulsionamento de conteúdos ilícitos, Clara Keller e Laura Mendes destacam que "a desmonetização de conteúdos não tem protagonismo em iniciativas regulatórias recentes, como é o caso das obrigações de transparência e devido processo, por exemplo". (KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 82).

Para os fins deste trabalho de pesquisa, que estuda especificamente os desafios relacionados à moderação de conteúdo, a regulação deve ser capaz de criar incentivos para que as plataformas se tornem instituições confiáveis e responsáveis no cumprimento dos valores centrais da liberdade de expressão – democracia política e cultural e disseminação de conhecimento – no contexto da curadoria do conteúdo *online*. A regulação é necessária para estimular, portanto, a criação e manutenção de uma esfera pública digital saudável, funcional e eficaz.

1.6.2. Buscar soluções para o problema de legitimidade decisória na moderação de conteúdo feita pelas plataformas digitais

Os modelos de regulação que estabelecem normas procedimentais para garantir transparência à moderação de conteúdo nas redes sociais têm se tornado cada vez mais relevantes. Afinal, normas procedimentais são fundamentais para a solução do atual problema de legitimidade das decisões tomadas pelas plataformas digitais na definição dos padrões de comunicação aceitos nas redes sociais e na própria moderação de conteúdo.[138] Algumas estratégias e medidas proativas têm sido pensadas para a solução do problema de legitimidade decisória, envolvendo, principalmente, deveres de notificação e devido processo na moderação de conteúdo, ou obrigações de transparência voltadas a diversas práticas.[139]

Historicamente, as plataformas digitais jamais tiveram incentivos para o compartilhamento de informações detalhadas e de forma transparente sobre a moderação de conteúdo. A reunião de informação sobre os procedimentos de moderação de conteúdo é uma tarefa complexa e trabalhosa, que pode envolver diversas ferramentas tecnológicas e equipes internas. Outro fator que inibe maior transparência decorre da divulgação ao público de qualquer erro, que pode ser utilizado contra as redes sociais nos tribunais ou pela imprensa.[140]

O problema se concentra atualmente na ausência de incentivos econômicos para que as redes sociais desempenhem o papel de moderadoras de conteúdo de forma confiável e responsável e promovam a curadoria de conteúdo que propicie a

---

[138] *Ibidem*, p. 75.
[139] *Ibidem*, p. 76.
[140] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation…*, pp. 4-5.

disseminação de conhecimento. Nesse contexto, as redes sociais acabam adotando políticas e práticas que prejudicam a saúde da esfera pública digital.[141]

Existem benefícios de longo prazo, no entanto, na adaptação do atual modelo – inclusive no contexto brasileiro – para garantir maior transparência à moderação de conteúdo, tanto para a sociedade quanto para as próprias empresas. Algumas plataformas passaram, inclusive, a divulgar periodicamente ao público relatórios de transparência sobre moderação de conteúdo, em virtude do reconhecimento desse problema e, ainda, do aumento das pressões por parte da sociedade civil e da academia.[142]

Qualquer debate sobre regulação da moderação de conteúdo das redes sociais deve passar, portanto, pela compreensão da importância de se garantir transparência ao modelo de curadoria de conteúdo *online*, sem se distanciar da complexidade das operações e das reais capacidades das plataformas digitais para o desempenho da moderação de conteúdo.

1.7. Regulação das redes sociais e sua tipologia

A doutrina afirma que a evolução da regulação da *internet* passou por uma primeira etapa – em torno de 1995 – em que a regulação foi desincentivada a favor da inovação. Nesse período, havia a divisão entre os veículos tradicionais de mídia, que haviam migrado ao ambiente virtual, e os usuários comuns. Com a evolução das redes sociais, em 2010, começam a surgir as plataformas digitais, que não apenas indexavam os conteúdos, como também os moderavam. A partir de então, o conteúdo de qualquer usuário – e não apenas os veículos de comunicação – tinha a possibilidade de viralizar, com o impulsionamento por algoritmos, sem qualquer comprometimento com o exercício responsável da liberdade de comunicação. O novo cenário tinha a finalidade de aumentar o engajamento dos usuários e, por extensão, incrementar a indústria publicitária. Surge, com isso, uma postura crítica em torno das redes sociais e um interesse gradual e contínuo em regulá-las.[143]

---

[141] BALKIN, Jack M. *How to Regulate (and Not Regulate) Social Media*…, p. 72.
[142] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*…, p. 5.
[143] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*. Belo Horizonte: Fórum, 2022, pp. 217-218.

De acordo com a doutrina, a internet é sustentada por três pilares essenciais que não apenas viabilizam sua existência, mas também definem sua funcionalidade e propósito: *(i)* a infraestrutura; *(ii)* o código; e *(iii)* o conteúdo.[144] A infraestrutura consiste nos elementos físicos e lógicos que sustentam a internet, como cabos, servidores, provedores de acesso, redes e *data centers*. Ela é a base técnica que torna possível a conexão global. O código, por sua vez, refere-se aos protocolos, *softwares* e padrões que determinam como a internet funciona e como dispositivos e sistemas se comunicam entre si. Já o conteúdo envolve o que é publicado, compartilhado ou consumido na rede, como textos, imagens, vídeos etc.

A regulação atua de maneira distinta sobre cada um desses aspectos. Na *infraestrutura*, as normas garantem acesso universal, neutralidade da rede, segurança e qualidade do serviço, assegurando que todos possam utilizar a internet de forma eficiente e equitativa. No *código*, as regras promovem a interoperabilidade entre sistemas, protegem a privacidade por *design*, regulam o uso de criptografia e impõem limites técnicos que orientam ou restringem certos comportamentos. Por fim, a regulação do *conteúdo* trata de temas sensíveis, como discurso de ódio, desinformação, direitos autorais, liberdade de expressão e a responsabilidade das plataformas.[145] Em vista disso, a regulação das redes sociais seria um espinhoso espectro da regulação (mais restrita) da *internet*, que exige a constituição de normas não apenas para regular a divulgação do *conteúdo* na *internet*, como também o próprio *código*.[146]

Considerando que foi respondido acima *o que* exatamente se pretende regular (redes sociais) e *porque* é preciso regular (liberdade de expressão e pluralismo); o passo seguinte é entender *como* (arquitetura institucional) a regulação deve acontecer.

Antes de discorrer sobre as regras constantes desse novo regime jurídico, no entanto, é necessário entender quais atores detêm a capacidade – técnica e legal – para prescrever comportamentos e competências nas redes sociais. Nesse sentido,

---

[144] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 43-104, p. 56.

[145] A discussão em torno do constitucionalismo digital e a aplicação horizontal de direitos fundamentais será devidamente tratada nos capítulos 2 e 3.

[146] O presente trabalho tem como enfoque apenas a regulação do conteúdo nas redes sociais, merecendo uma dissertação a parte para tratar da regulação a partir dos códigos que constituem a internet e as redes sociais.

Domingos Soares Farinho descreve *tipos ideais de regulação* a partir da *fonte normativa* que constituiriam essas regras, dividindo-a do seguinte modo: *(i)* a regulação privada ou autorregulação; *(ii)* a regulação pública ou heterorregulação; e *(iii)* a corregulação ou autorregulação regulada.[147]

1.7.1. Autorregulação ou regulação privada

A autorregulação ou regulação privada das redes sociais envolve o exercício da autonomia privada e do direito de propriedade das plataformas (CF, arts. 5º, II e XXII).[148] A partir do livre exercício da atividade econômica, as plataformas determinam os princípios e regras da participação em sua comunidade com os termos de uso (CF, artigo 170, parágrafo único). Nesses contratos de adesão, em que estão de um lado os fornecedores de serviço (redes sociais) e do outro consumidores (usuários),[149] são desenvolvidos parâmetros para determinar/tutelar o que seria condizente – ou não – com o exercício da liberdade de expressão, considerando a comum colisão com outros direitos individuais e coletivos (*v.g.*, honra, nome, integridade psicofísica, privacidade, proteção de dados pessoais,

[147] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 43-104, p. 56.

[148] BOBBIO, Noberto. Teoria do ordenamento jurídico. 6ª ed. Maria Celeste C. J. Santos (trad.). Cláudio De Cicco (rev. téc.). Brasília: Editora Universidade de Brasília, 1995, p. 40: "Outra fonte de normas de um ordenamento jurídico é o poder atribuído aos particulares de regular, mediante atos voluntários, os próprios interesses: trata-se do chamado poder de negociação. O enquadramento dessa fonte na classe das fontes reconhecidas ou na das fontes delegadas é menos nítido. Se se coloca em destaque a autonomia privada, entendida como capacidade dos particulares de dar normas a si próprios numa certa esfera de interesses, e se considerarmos os particulares como constituintes de um ordenamento jurídico menor, absorvido pelo ordenamento estatal, essa vasta fonte de normas jurídicas é concebida de preferência como produtora independente de regras de conduta, que são aceitas pelo Estado."

[149] KONDER, Carlos Nelson de Paula; SOUZA, Amanda Guimarães Cordeiro de. Onerosidade do acesso às redes sociais. *Revista de Direito do Consumidor*, ano 28, vol. 121, jan.-fev./2019, pp. 185-212: "(...) a doutrina consumerista já percebera que diversos contratos não geram custos para o consumidor, mas mesmo assim possibilitam ao fornecedor obter vantagens económicas, remunerando-o indiretamente. Serviços desse tipo são hoje submetidos ao Código de Defesa do Consumidor. A novidade é que o aludido entendimento da doutrina, e também da jurisprudência, vem sendo aplicado, da mesma forma, aos contratos de prestação de serviços pretensamente gratuitos pela Internet, renovando-se a menção aos ganhos indiretos dos fornecedores. As vezes ditos ganhos indiretos do fornecedor – no caso, de redes sociais – são explicados pela renda de publicidade. Entretanto, a grande riqueza da atualidade, e que os consumidores das redes sociais oferecem, consiste nos dados pessoais revelados enquanto o usuário acessa e produz conteúdo e interage com outros perfis naquela plataforma, ensinando os fornecedores de redes sociais sobre seus gostos e hábitos. A conclusão, assim, em uma perspectiva funcional, é que esses contratos são qualificados como onerosos e, por conseguinte, sujeitos à aplicação do Código de Defesa do Consumidor."

saúde pública, jornalística, integridade do sistema eleitoral ou financeiro etc. – CF, artigos 5º, inciso IV e IX, e artigo 220).[150]

A discussão, até aqui, é bastante familiar aos juristas. A dificuldade emerge quando se verifica que as normas privada que regulam a liberdade de expressão (e, por extensão, as colisões dela advindas) concretizam-se por um regime jurídico libertário – como é o estadunidense[151] – e é executado por *códigos*, que, segundo Lawrence Lessig, são a própria lei no ambiente digital.[152]

Embora se consiga avaliar, por pesquisas externas, o impacto dos algoritmos na moderação de conteúdo ou remoção de contas,[153] é razoável entender que tais elementos são, ainda, uma verdadeira caixa-preta à sociedade, que, diluída em argumentos de propriedade intelectual, não demonstram seu verdadeiro impacto ou os interesses privados que perseguem. A autorregulação, seus fundamentos e métodos serão esmiuçados, pormenorizadamente, no capítulo 2 *infra*.

1.7.2. Heterorregulação ou regulação pública

Conforme desenvolvido acima, a divisão tripartite tem o modesto objetivo de indicar a *fonte normativa* da regulação aplicável: *(i)* privada; *(ii)* pública; ou *(iii)*

---

[150] WIELSCH, Dan. Os ordenamentos das redes – Termos e condições de uso – Código – Padrões de comunidade. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 105-133, p. 117: "Em razão de sua formulação unilateral e da estrutura supraindividual de relações de uso, têm, na prática, o caráter de normas gerais abstratas. A vinculatividade jurídica na relação horizontal não pode resultar de um consenso transacional alcançado entre as partes sobre todos os conteúdos relevantes (cf. art. 154 e ss. do Código Civil alemão), mas deriva, sim, de uma concordância global com a validade, à qual não é possível associar nenhum tipo de suposição de correção. Consequentemente, até por razões de legitimidade da criação de normas por particulares, faz-se necessário um controle legal da correção."

[151] No caso *Murthy, Surgeon General*, et al. v. *Missouri et al.*, a Suprema Corte dos Estados Unidos analisou se autoridades federais violaram a Primeira Emenda ao interagir com plataformas de redes sociais para combater desinformação. Estados como Missouri alegaram que o governo pressionou empresas privadas para moderar conteúdos, configurando censura governamental indireta. A Corte teve de ponderar os limites do poder público no combate à desinformação em plataformas privadas, balanceando o papel do governo na proteção da saúde pública com os direitos constitucionais dos cidadãos. A decisão destacou que o governo pode colaborar com plataformas para promover interesses públicos legítimos, como combater informações falsas sobre saúde, desde que tal colaboração não se transforme em coerção ou censura. O Tribunal enfatizou que interações entre autoridades e empresas devem respeitar a autonomia das plataformas, preservando os direitos de expressão dos usuários e evitando interferências diretas ou indiretas nos conteúdos. Esse precedente reafirma a necessidade de manter a liberdade de expressão mesmo diante de desafios contemporâneos, como a propagação de desinformação *online* (SUPREMA CORTE DOS ESTADOS UNIDOS DA AMERICA. *Murthy, Surgeon General, et al. v. Missouri et al.* 06 jun. 2024. Disponível em: <www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf.> Acesso em 13 jan. 2025).

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

[153] Remeta-se ao subcapítulo 1.6 *supra*.

privada-pública. No Ocidente, ao que se tem notícia, nenhum país constituiu uma agência ou empresa pública para desenvolver e disponibilizar uma rede social à sua população, apesar da possibilidade. Por causa disso, nenhum Estado ocidental adotou uma regulação *puramente* pública para o exercício da liberdade de expressão nas redes sociais.

Apesar disso, diversos países implementaram restrições e/ou bloqueios a plataformas como *Instagram*, *Google* e *YouTube*, motivados por razões políticas, de segurança ou controle de informação. Na China, desde 2009, o país bloqueia plataformas como *Instagram* e *Google*, visando controlar o fluxo de informações e manter a estabilidade social, extinguindo manifestações pró-democracia, como apontam os jornais.[154] Em 2022, durante o conflito com a Ucrânia, a Rússia bloqueou o acesso ao *Instagram*, acusando a plataforma de permitir "que usuários das redes sociais em alguns países defendam atos de violência contra russos".[155]

Diante disso, é preciso mencionar que mesmo países reconhecidamente interventores na liberdade de expressão de seus cidadãos, não se consegue desfazer da organização e investimentos privados necessários ao desenvolvimento de uma rede social de sucesso; os países adotam alternativas locais. Na China, plataformas como *WeChat*, que combina mensagens e pagamentos,[156] e *Weibo*, semelhante ao *Twitter* (atual rede "X"),[157] dominam o mercado. Outras opções populares incluem *Douyin*, versão chinesa do *TikTok*,[158] e *Baidu Tieba*, um fórum temático. Ainda assim, a China tem empreendido uma investigação minuciosa acerca de alegadas violações ao regime jurídico chinês, que comprometeria – em tese – a segurança

---

[154] REAUTERS. Instagram é bloqueado na China em meio a protestos em Hong Kong. *G1*, 29 set. 2014. Disponível em: <https://g1.globo.com/tecnologia/noticia/2014/09/instagram-e-bloqueado-na-china-em-meio-protestos-em-hong-kong.html>. Acesso em 13 jan. 2025.

[155] G1. Instagram restrito na Rússia: entenda a importância da rede social para o país de Putin. G1, 12 mar. 2022. Disponível em: <https://g1.globo.com/tecnologia/noticia/2022/03/12/instagram-restrito-na-russia-entenda-a-importancia-da-rede-social-para-o-pais-de-putin.ghtml>. Acesso em 13 jan. 2025.

[156] SILVEIRA, Janaína. WeChat: o app faz tudo que mudou a vida dos chineses. Veja, 23 nov. 2018. Disponível em: <https://veja.abril.com.br/mundo/wechat-o-app-faz-tudo-que-mudou-a-vida-dos-chineses>. Acesso em 13 jan. 2025.

[157] FORBES BRASIL. Weibo, o Twitter chinês, já é mais valioso que o original. *Forbes*, 12 out. 2016. Disponível em: <https://forbes.com.br/negocios/2016/10/weibo-o-twitter-chines-ja-e-mais-valioso-que-o-original/>. Acesso em 13 jan. 2025.

[158] CUETO, José Carlos. Rússia restringe acesso ao Facebook e Twitter após ataques à Ucrânia. *BBC News Brasil*, 26 fev. 2022. Disponível em: <https://www.bbc.com/portuguese/articles/cq5zydp59j7o>. Acesso em 13 jan. 2025.

nacional.[159] Como um gênio que, uma vez liberto, não volta a lâmpada, as empresas de tecnologia parecem integrar a socialidade moderna, tornando inimaginável – até o momento – um sistema puramente estatal.

Na heterorregulação ou regulação estatal, a fonte normativa é o Estado e decorre de seus respectivos processos legislativos e/ou administrativos. De acordo com a ordem constitucional brasileira, a União tem competência exclusiva para legislar sobre direito civil, comercial, proteção de dados e bases da comunicação (CF, artigo 22, I). Essa competência garante uniformidade na regulação das plataformas digitais, assegurando proteção aos usuários e harmonização com os direitos constitucionais.

Além disso, a proteção de direitos fundamentais, como a privacidade e a intimidade (CF, artigo 5º, X) e a defesa do consumidor (CF, art. 5º, XXXII), também fundamenta a regulação das redes sociais. O artigo 170, que trata da ordem econômica, exige que o funcionamento de plataformas respeite princípios como a soberania nacional (inc. I), livre concorrência (inc. IV), defesa dos consumidores (inc. V) e meio ambiente (inc. VI), consolidando a legitimidade da União para estabelecer normas que promovam transparência e equilíbrio nas relações digitais.

Assim, a regulação pública das relações entre redes sociais e seus usuários deve ser limitada à garantia de que não imponha restrições diretas e desproporcionais à livre iniciativa e à autonomia privada das redes sociais. A legitimidade de uma intervenção pública parece aceitável em situações de prevenção, correção ou repressão de danos causados nas redes sociais, quando as plataformas digitais falharem nesse papel. Dessa forma, a regulação pública dessas relações deve seguir um princípio de subsidiariedade, alinhando-se, como será proposto, aos modelos de corregulação ou autorregulação regulada.[160]

1.7.3. Autorregulação regulada ou corregulação

Devido à ausência de compromisso com interesses públicos,[161] as redes sociais têm sido alvo de preocupação tanto para Estados, relaciona à integridade de

[159] BBC. China's WeChat, Weibo and Baidu under investigation. *BBC News*, publicado em 11 ago. 2017. Disponível em: <https://www.bbc.com/news/world-asia-china-40896235>. Acesso em 13 jan. 2025.
[160] O desafio da regulação pública será abordado no Capítulo 3 *infra*.
[161] SAAD, Beth; MALAR, João Pedro. Mudanças da Meta ilustram a nova e perigosa era das redes sociais. *Folha de S. Paulo*, São Paulo, 12 jan. 2025. Disponível em:

seus ordenamentos jurídicos, quanto para o mercado, que aponta o inevitável desgaste das marcas com a associação de redes sociais tóxicas.[162] O Brasil já demonstra o desgaste gerado pela exclusiva autorregulação das redes,[163] reforçando uma coalizão internacional para reduzir a discricionariedade das redes sociais.[164]

Por causa disso, torna-se cada vez mais comum a inclinação no Brasil de regular a autorregulação já existente em nosso mercado – ou seja, construir um sistema de autorregulação regulada ou corregulação.[165] Nesse cenário, existe a combinação da fonte normativa privada e pública com a finalidade de adequar a expertise das empresas de tecnologia e comunicação com os interesses públicos necessário à concretização da ordem constitucional.[166] O movimento brasileiro, é

<https://www1.folha.uol.com.br/opiniao/2025/01/mudancas-da-meta-ilustram-a-nova-e-perigosa-era-das-redes-sociais.shtml>. Acesso em 13 jan. 2025.

[162] MULHOLLAND, Isabella. Quem anunciará onde fake news e conteúdo tóxico correm soltos? *Meio&Mensagem*, 14 dez. 2023. Disponível em: <https://www.meioemensagem.com.br/opiniao/quem-anunciara-onde-fake-news-e-conteudo-toxico-correm-soltos>. Acesso em 13 jan. 2025.

[163] HOLANDA, Marianna. Lula convoca reunião sobre Meta e diz que um cidadão não pode ferir soberania da nação. *Folha de S. Paulo*, São Paulo, 12 jan. 2025. Disponível em: <https://www1.folha.uol.com.br/poder/2025/01/lula-convoca-reuniao-sobre-meta-e-diz-que-um-cidadao-nao-pode-ferir-soberania-da-nacao.shtml>. Acesso em 13 jan. 2025.

[164] CARLUCCI, Manoela. Por telefone, Lula e Macron conversam sobre decisão da Meta. *CNN Brasil*, 12 jan. 2025. Disponível em: <https://www.cnnbrasil.com.br/politica/por-telefone-lula-e-macron-conversam-sobre-decisao-da-meta/>. Acesso em 13 jan. 2025.

[165] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 66: "Diante disso, apresenta-se para o Brasil o desafio de estabelecer uma regulação de plataformas estrutural, que não apenas enderece expressões de uma crise de erosão democrática, mas que garanta ao país o aparato institucional necessário à proteção de direitos e promoção da inovação em uma sociedade que se torna cada vez mais digitalizada. Tal política requer, além de uma regulação horizontal das plataformas digitais, também legislações verticais de áreas específicas, incluindo medidas como as de cunhos 'econômicos e concorrenciais, do trabalho, de proteção à criança, inclusão digital e de promoção do jornalismo de qualidade e do conhecimento'."

[166] MARANHÃO, Juliano; CAMPOS, Ricardo. Fake News e autorregulação regulada das redes sociais no Brasil: fundamentos constitucionais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 341-355, pp. 345-346: "Uma forma moderna de lidar com a crescente incerteza do ponto de vista regulatório encontra-se no instituto da autorregulação regulada. Este procura trabalhar no limite entre duas tradicionais formas de regulação: autorregulação e regulação por um terceiro, normalmente o Estado ('*Fremdregulierung*'). Por um lado, a autorregulação tem a vantagem da eficiência pela disposição do conhecimento interno e dinâmica de constante revisão de conceitos. Por outro, tem a desvantagem de não necessariamente perseguir interesses e valores públicos. Já a regulação por terceiro ('*Fremdregulierung*') tem a vantagem de poder ser implementada por coerção em nome do interesse público e a desvantagem de, em ambientes dinâmicos, não dispor de conhecimento suficiente para lograr êxito na persecução do objetivo. A autorregulação regulada oferece outra e nova possibilidade de lidar com as incertezas, ao conciliar vantagens das duas abordagens alternativas: a regulação versus a autorregulação. Foca-se no importante momento de auto-organização conforme expertise e dinâmica própria da indústria, estimulando-se, porém, alguns parâmetros gerais de interesses públicos caros ao Estado e à sociedade. Nesse sentido, a autorregulação regulada consegue 'induzir' o setor privado a contribuir para o cumprimento de tarefas públicas. Essa forma de regulação pode lidar melhor com uma sociedade que cada vez mais

claro, não está isolado, encontrando eco na Alemanha (*NetzDG*) e na Inglaterra (*Online Safety Bill*), conforme mencionado acima.

A regulação das redes sociais apresenta uma natureza *policêntrica*, refletida na dispersão de metodologias ponderadoras que emergem de distintas percepções sobre os interesses envolvidos.[167] Essa característica se evidencia como um fundamento corregulatório, alinhado à natureza dos agentes reguladores, estabelece uma ponderação jurídica mais consentânea aos interesses público e privado.

A contramão disso seria persistir em um modelo regulatório (como a estadunidense) que seria exportável a todos os países.[168_169] Assim, a corregulação impõe-se como uma resposta mais adequada do sistema jurídico às complexidades inerentes às redes sociais, corrigindo, inclusive, a tendência de sobrecarregar o poder judicial como um regulador da autorregulação ou da regulação privada.

---

se locomove, e se distancia, de uma sociedade centrada em organizações, conseguindo absorver melhor as incertezas e construir parâmetros melhores de eficiência na regulação."

[167] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., pp. 43-104, p. 56: "As redes sociais, como expressões da internet, compartilham suas características e, em alguns casos, as exibem de forma mais clara e evidente do que outros domínios da rede. Sendo entidades privadas que exercem uma auto-organização de suas atividades, mas que estão sujeitas ao controle público pelos Estados onde disponibilizam suas plataformas, as redes sociais tornam-se objeto de uma regulação que a doutrina designa como policêntrica. Essa abordagem não pode ser reduzida a uma única fonte, seja ela privada ou pública. Essa evidência jurídica estruturante configura-se como fundadora de um dever de corregulação, conforme será demonstrado. Tal dever é essencial para evitar violações das regras que sustentam a autonomia privada ou a prossecução do interesse público, destacando a complexidade e a necessidade de colaboração entre diferentes atores para a regulação das redes sociais."

[168] A preocupação se torna cada vez mais concreta com o avanço da desconstituição dos checadores de fato e outros filtros de moderação de conteúdo por Mark Zuckerberg. Recentemente, o empresário afirmou que se estaria formando uma crescente institucionalização da censura em contraste com "[o]s EUA [que] têm as mais fortes proteções constitucionais para a liberdade de expressão no mundo". Ao invés de assumir a natural pluralidade política e cultura, o empresário se limitou a mencionar a ordem constitucional de seu país, sustentando que as redes sociais do grupo Meta estariam "voltando às suas raízes" (UOL. Zuckerberg diz que Meta vai acabar com checagem de fatos, cita censura e manda recado ao STF. *UOL*, publicado em 7 jan. 2025. Disponível em: <https://www.youtube.com/watch?v=nJQt3DLQqQ0>. Acesso em 13 jan. 2025). Nesse contexto, há de ressaltar, ainda, a proposta de descentralização da moderação de conteúdo, por meio de "notas da comunidade", que permitem que os próprios usuários das plataformas possam agregar contexto ou mesmo refutar o conteúdo de uma publicação, em relação às quais o ITS Rio alerta para a necessidade "de mais estudos para que possa ser devidamente avaliada" (ITS RIO. *7 reflexões para o futuro do debate sobre moderação de conteúdo em plataformas digitais*. Publicado em 19 mar. 2025. Disponível em: https://itsrio.org/pt/publicacoes/moderacao-conteudo-plataformas-digitais-its-rio/. Acesso em 7 abr. 2025).

[169] Após a Advocacia Geral da União (AGU) notificar a Meta acerca da declaração sobre a mudança de política de moderação de conteúdo, houve um abrandamento do tom: "A Meta desde já esclarece que, no momento, está encerrando seu Programa de Verificação de Fatos independente apenas nos Estados Unidos, onde testaremos e aprimoraremos as Notas da Comunidade antes de dar início a qualquer expansão para outros países" (BRASIL. Advocacia-Geral da União (AGU). *Carta para resposta à notificação extrajudicial*. Disponível em: <https://www.gov.br/agu/pt-br/nota-agu-recebe-manifestacao-da-meta/Cartapararespostaanotificacaoextrajudicial_13.1.20251.pdf>. Acesso em 17 jan. 2025).

2. AUTORREGULAÇÃO: LEGITIMIDADE, LIMITES E DESAFIOS DA MODERAÇÃO DE CONTEÚDO NAS REDES SOCIAIS COM BASE NA APLICAÇÃO DOS TERMOS DE USO

No ordenamento jurídico brasileiro, as redes sociais avaliam, por seus próprios critérios previamente estabelecidos nos Termos de Uso, o conteúdo que deve ser moderado e as contas de usuários que devem ser suspensas ou removidas, cabendo ao Poder Judiciário, pela regra geral, a aplicação da responsabilidade civil em caso de desobediência de ordem judicial prévia (MCI, artigo 19).

Conforme analisado no capítulo 1 *supra*, a autorregulação se tornou problemática quando a circulação de *fake news* passou a gerar convulsões institucionais graves (*v.g.*, movimento antivacina no Covid-19). Nesse contexto, a constitucionalidade do artigo 19 do MCI foi examinada pelo STF, que, pelos votos dos ministros disponíveis até o momento, buscou criar um sistema de corregulação, em que as redes sociais passam a deter determinados "deveres de cuidado" (*duty of care*).[170] Uma tarefa ingrata, sem dúvidas, considerado os limitadíssimos escopos das demandas judiciais e, principalmente, os apelos à regulação estrutural do tema realizados nos votos pelos próprios ministros da Corte.[171]

Embora pareça natural a discussão, de um lado, em torno ponderação de interesses públicos e, de outro, o exercício da liberdade editorial na moderação de seu conteúdo, é razoável considerar que ainda persiste uma lacuna em torno de quais institutos privados estariam efetivamente em jogo. A questão é central. Como a qualificação jurídica é essencial para se identificar qual o regime jurídico aplicável,

[170] O tema foi discutido no subcapítulo 1.5.2 *supra* em comentário ao julgamento dos Recursos Extraordinários 1.037.396/SP (Tema 987) e 1.057.258/MG (Tema 533).

[171] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., p. 68. De acordo com as autoras, ainda que exista um regime de responsabilidade de intermediários judicial e ele seja reconhecido como boa prática regulatória, existem limitações significativas. Primeiro, o poder judiciário só atua quando acionado, restringindo o acesso ao escrutínio público a quem tem condições materiais e subjetivas de recorrer à justiça. Além disso, esse modelo delega às plataformas privadas grande parte das decisões sobre liberdade de expressão *online*, expondo limitações do judiciário, como seletividade e falta de expertise técnica, especialmente em um contexto complexo como o das redes digitais. Por fim, as redes sociais concentram crescente poder econômico e político, afetando profundamente instituições e práticas sociais. Assim, regulações mais abrangentes e estruturais são necessárias para equilibrar liberdade econômica e direitos fundamentais. Por fim, a regra geral aplicada a diferentes tipos de conteúdos infratores, como incitação à violência e desinformação, carece de critérios proporcionais à gravidade e aos danos potenciais, destacando a necessidade de abordagens mais específicas e eficazes.

é necessário considerar o que seria o exercício da moderação de conteúdo por parte das redes sociais.

No panorama atual, *Meta* (*Facebook* e *Instagram*), *Google* (*Youtube*) e *X* são pressionadas a moderar conteúdos ilícitos de terceiros em suas redes sociais sem a necessidade prévia de manifestação judicial, a despeito da ausência de previsão legal (sistema de autorregulação).[172]

Nesse contexto, as redes sociais funcionam como verdadeiro anteparo à inundação de casos que atingiria o próprio Poder Judiciário – caso fossem desativados os controles de conteúdo automáticos – e moderadores do próprio debate público, no contexto do combate às *fake news*.[173] Contudo, existe aí uma aparente antinomia: onde estaria a inafastabilidade jurisdicional necessária para se excluir o conteúdo de um cidadão ou autoridade pública? É possível que uma empresa privada "de ofício" possa alterar o discurso público sem nenhuma necessidade de percorrer um caminho judicial?

Embora a jurisprudência já tenha decidido, sem poder vinculativo, que as redes sociais têm poderes de exercer sua liberdade editorial desde que em consonância com o ordenamento jurídico brasileiro,[174_175] os "prejudicados" continuam apontando a falta de controle no exercício da autorregulação das redes sociais. Existe essa inevitável tormenta em torno dessas situações jurídicas porque, ao que parece, a capacidade de moderar o conteúdo é exercida pelas plataformas digitais a partir da autossatisfação do direito à sua liberdade editorial, sem a necessidade de se recorrer à tutela jurisdicional, como um possível mecanismo de *autotutela* das redes sociais.

---

[172] De acordo com Elon Musk, a rede social X recebeu, desde a instauração do referido inquérito, entre 2019 e 2024, 88 (oitenta e oito) e 29 (vinte e nove) ordens do STF e do Superior Tribunal Eleitoral, respectivamente. Em manifestação oficial, Elon Musk afirmou que os principais alvos de bloqueio eram membros do Congresso e reconhecidos jornalistas brasileiros. Ver, para mais informações, FALCÃO, Márcio. *Moraes dá cinco dias para o X explicar lives de contas bloqueadas pela Justiça*. TV Globo, publicado em 22 abr. 2024. Disponível em: <https://g1.globo.com/politica/noticia/2024/04/22/moraes-da-cinco-dias-para-o-x-explicar-lives-de-contas-bloqueadas-pela-justica.ghtml>. Acesso em 17 jan. 2024.

[173] Remeta-se ao capítulo 1 *supra*.

[174] BARROSO, Luís Roberto; BARROSO, Luna van Brussel. Democracia, mídias sociais e liberdade de expressão: ódio, mentiras e a busca da verdade possível. *Direitos Fundamentais & Justiça*, Belo Horizonte, ano 17, n. 49, p. 285-311, jul./dez. 2023.

[175] Houve manifestação amplamente fundamentada em torno da moderação de conteúdo de *fake news* durante a pandemia de Covid-19 (BRASIL. Superior Tribunal de Justiça. REsp 2139749/SP, Rel. Min. Ricardo Villas Bôas Cueva, 3ª Turma, j. 27 dez. 2024).

Um poder cujo principal requisito é exatamente a previsão legislativa expressa,[176] que existe, por exemplo, para combater o esbulho de terras (desforço possessório – CC, art. 1.210), a venda de ações do acionista remisso em bolsa de valores (Lei 6.404/1976, art. 107, II), a venda do bem objeto da propriedade fiduciária no âmbito do mercado de capital (Lei 4728/1965, art. 66-B, §3°), a apropriação da caução locatícia (Lei 8245/1991, art. 37, I), a apropriação das arras (CC, art. 418 – autotutela da cláusula penal) etc., mas não especificamente para moderar o conteúdo das redes sociais, exceto pelo regime legal de responsabilidade que restringe a responsabilização das plataformas aos casos em que tenha havido descumprimento da obrigação de executar determinada ordem judicial. A diferença reside na atipicidade inerente aos termos de uso.[177]

---

[176] "A resistência à análise de medidas que fujam da supervisão do Poder Judiciário é amplamente criticada, por exemplo, nos defensores do pacto comissório. Após citar exemplos de execução extrajudicial, Carlos Edison do Rêgo Monteiro Filho afirmou que "o ordenamento jurídico convive harmonicamente com instrumentos de execução de direitos à revelia do Poder Judiciário. Nota-se que o legislador ordinário buscou adequar a indigitada autotutela às pretensões surgidas no seio social ao longo do século XX, que urgem por mecanismos mais céleres (e eficazes) de resolução de litígios, dada a complexidade crescente das relações contratuais, a tornar fugaz o tráfego econômico e jurídico, em contraste com a morosidade dos tribunais, que, na contramão, se encontram abarrotados de processos os mais diversos. Não se trata, portanto, de reacender a chama do exercício das próprias razões mas de tutelar o anseio generalizado por sistema jurídico de atuação prática efetivamente dinâmica, alinhado ao contexto histórico social no qual se insere. Ponderadas as razões, de parte a parte, deduzidas supra, percebe-se a insuficiência da inderrogabilidade do procedimento judicial fundamento apto, por si, a proibir o pacto comissório." (MONTEIRO FILHO, Carlos Edison do Rêgo. *Pacto comissório e pacto marciano no sistema brasileira de garantias*. Rio de Janeiro: Editora Processo, 2017, p. 43-44). Isso também foi alertado, outrora, por Aline de Miranda Valverde Terra, ao afirmar, em defesa da cláusula resolutiva expressa, "[a] sociedade tem exigido um novo desenho da planta jurisdicional, apoiada na constatação, hoje inequívoca, de que a tutela justa e tempestiva dos seus direitos não constitui um monopólio do Poder Judiciário. O fenômeno crescente da desjudicialização eleva entre nós a importância dos instrumentos de autotutela, sem que disso decorra qualquer tipo de violação, *a priori*, ao princípio da inafastabilidade da jurisdição." (TERRA, Aline de Miranda Valverde Terra. Inafastabilidade da jurisdição e autotutela: o exemplo da cláusula resolutiva expressa. *Revista Eletrônica de Direito Processual*, Rio de Janeiro, ano 13, vol. 20, n. 3, p. 1-19, set./ dez. 2019).

[177] CABRAL, Antônio do Passo. Repensando a autotutela: conceito e limites no direito brasileiro. *Revista de Processo*, abril, 2024, vol. 350, ano 49, pp. 21-47, p. 32: "Embora, como dito, a autotutela tenha ficado à margem dos estudos jurídicos durante muito tempo, o tema tem retornado à reflexão dos juristas. E, nos últimos anos, mecanismos de autotutela começaram a ser revistos como uma alternativa autocompositiva ao sistema contencioso judiciário. Pelos custos e tempo envolvidos em um processo judicial, buscam-se mais e mais métodos confiáveis, seguros, rápidos e eficientes para as partes reagirem ao ilícito e protegerem seus direitos subjetivos. E uma maneira eficaz para as partes de um contrato escaparem da burocrática justiça estatal é o desenvolvimento de *formas consensuais de autotutela*. Note-se bem: aqui, a autotutela também é de exercício unilateral, mas a previsão normativa que autoriza é uma regra contratual. Portanto, na origem, a autotutela é permitida pela vontade consensual das partes. À luz dessas novas formas convencionais, a autotutela, que já possuía previsão legislativa no direito brasileiro, passou também a ter fonte negocial. Isso ocorre quando as partes, por meio de um negócio jurídico, criam normas que permitem, diante de um conflito, a imposição de prestações contratuais de forma unilateral por um dos contratantes. É justamente por meio dessas possibilidades de instituição de autotutela via contrato que o tema ganha

Há, portanto, uma certa fratura em como se enxerga a autotutela no direito brasileiro. Embora parte majoritária da doutrina clássica defenda, em respeito à dogmática lançada nos idos do século XX, a autotutela como um remédio excepcional, que exige previsão legislativa expressa,[178] a prática tem indicado que a autotutela também pode ser necessária para assegurar a tutela de interesses legítimos – como o combate às *fake news*, aos discursos de ódio e à violência nas redes sociais – por meio autoexecutoriedade das sanções previstas nos termos de uso das redes sociais.[179]

A doutrina contemporânea tem se esforçado, nas últimas décadas, para conceber a autotutela antes como "parte da ordem geral da tutela dos direitos" do que exceção às regras do ordenamento jurídico.[180] Com isso, autores argumentam, de forma técnica, a necessidade de incorporar institutos qualificados como excepcionais como parte de um esquema geral, como tem sido a aproximação do direito de retenção à exceção do contrato não cumprido.[181]

Como o presente trabalho de pesquisa se debruçará mais especificamente sobre o sistema da autorregulação regulada para as redes sociais, é preciso demonstrar o fundamento jurídico de ambos. Nesse capítulo, pretende-se desenvolver, portanto, o fundamento jurídico, limites e funções da autorregulação.

Para isso, serão desenvolvidas as práticas aplicadas pelas empresas de tecnologia, para avaliar se essas práticas poderiam ser qualificadas dogmaticamente como sanções contratuais previstas nos termos de uso, que podem ser impostas pelas plataformas digitais, independentemente da prévia tutela jurisdicional,

---

relevância. A autonomia negocial permite que as partes estabeleçam mecanismos de autotutela, independentemente de previsão legal. Nesse sentido, é possível pensar em uma atipicidade das formas de autotutela, e é neste ponto que pode residir o futuro mais promissor do instituto nas sociedades contemporâneas."

[178] A relação entre o princípio da legalidade e a autotutela será desenvolvido no subcapítulo 2.1.1 *infra*.

[179] SALLES, Raquel Bellini de Oliveira. *Autotutela nas relações contratuais*. Rio de Janeiro: Editora Processo, 2019, pp. 36-37: "A observação é importante para desconstruir eventual entendimento de que os meios alternativos de solução de controvérsia, e a própria autotutela que se defende nesse trabalho, teriam a função de precípua de reduzir o inchaço da máquina judiciária. Mais do que isso, e antes disso, referidos mecanismos surgem e crescem a partir da necessidade de se reconhecer à autonomia privada possibilidades efetivas de atuação em defesa de interesses legítimos, assegurando-se um sistema mais amplo de tutela de direitos. O colapso da máquina judiciária é, assim, fator que contextualiza e impulsiona o movimento de desjudicialização, não seu fundamento."

[180] *Ibidem*, p. 79.

[181] Em tentativa de sistematizar o debate sobre o instituto da retenção, rejeitando sua alegada taxatividade, SILVA, Rodrigo da Guia. Notas sobre o cabimento do direito de retenção: desafios da autotutela no direito privado. *Civilistica.com*. Rio de Janeiro, a. 6, n. 2, 2017. Disponível em: <http://civilistica.com/notas-sobre-o-cabimento-do-direito-de-retencao/>. Acesso em 12 jan. 2024.

configurando, nesse sentido, uma possível nova forma de autotutela admitida pelo ordenamento jurídico brasileiro.

2.1. A aplicação da sanção contratual por autotutela? Fundamento, estrutura e função

A inafastabilidade da jurisdição garante aos cidadãos acesso à justiça, oferecendo tutela contra qualquer ameaça ou violação de direitos.[182] Em vista disso, o artigo 5º, XXXV, da Constituição Federal, assegura que "a lei não excluirá da apreciação do Poder Judiciário lesão ou ameaça a direito". Estabelece-se também, em prestígio aos economicamente mais vulneráveis, a gratuidade do acesso à justiça "aos que comprovarem insuficiência de recursos".[183]

A situação se complica, no entanto, quando, pela instantaneidade das relações socioeconômicas, a miríade de redes sociais e aplicativos de mensagens, contrastou com o conservadorismo do Poder Judiciário em atualizar sua atuação. Em exemplo singelo, as citações eletrônicas e a manutenção dos dados cadastrais atualizados de pessoas jurídicas de grande porte, indispensáveis hoje, foram formalmente aceitas com a vigência da Lei 14.195/2021, há menos de 5 anos, reduzindo a duração dos processos.

Nos últimos meses, o Presidente do Conselho Nacional de Justiça e do STF, Ministro Luis Roberto Barroso, também tem trabalhado para ampliar a utilização

---

[182] "Tais princípios procuraram refletir o novo direcionamento dos fins a que o processo modernamente se propõe como instrumento ético, acessível a todos, operoso, proporcional e útil do ponto de vista prático, a serviço do justo, e terão as seguintes denominações: princípio da acessibilidade, da operosidade, da utilidade e da proporcionalidade. (...) A acessibilidade pressupõe a existência de pessoas, em sentido lato (sujeitos de direito), capazes de estar em juízo, sem óbice de natureza financeira, desempenhando adequadamente o seu labor (manejando adequadamente os instrumentos legais judiciais e extrajudiciais existentes), de sorte a possibilitar, na prática, a efetivação dos direitos individuais e coletivos, que organizam uma determinada sociedade (...). Esse princípio [operosidade] significa que as pessoas, quaisquer que sejam elas, que participam direta ou indiretamente da atividade judicial ou extrajudicial, devem atuar da forma mais produtiva e laboriosa possível para assegurar o efetivo acesso à justiça. (...) E fundamental que o processo que o processo possa assegurar ao vencedor tudo aquilo que ele tem direito a receber, da forma mais rápida e proveitosa possível, com menor sacrifício para o vencido. A jurisdição ideal seria aquela que pudesse, no momento mesmo da violação, conceder, a quem tem razão, o direito material. (...) O julgador projeta e examina os possíveis resultados, as possíveis soluções, faz a comparação entre os interesses em jogo, e, finalmente, a opção, a escolha daquele interesse mais valioso, o que se harmoniza com os princípios e os fins que informam este ou aquele ramo do direito. Esta atividade retrata a utilização do princípio da formalidade". (CARNEIRO, Paulo Cezar Pinheiro. *Acesso à justiça:* juizados especiais cíveis e ação civil pública — uma nova sistematização da teoria geral do processo. Rio de Janeiro: Forense, 2000, p. 54).

[183] Constituição Federal, Art. 5º, LXXIV: "O Estado prestará assistência jurídica integral e gratuita aos que comprovarem insuficiência de recursos".

de inteligência artificial, incrementando a prestação do serviço jurisdicional.[184] Sob o título de racionalização do processo, esse lento movimento de atualização busca reduzir as críticas de ineficiência do Poder Judiciário na condução de seus procedimentos.

Contudo, em 31 de dezembro de 2023, havia um acervo de 83,8 milhões de processos em tramitação, incluindo suspensos, sobrestados e em arquivamento provisório. Nesse ano, foram recebidos três milhões de casos novos a mais do que em 2022. Apesar desse aumento, a alta produtividade atenuou o impacto, resultando em um saldo de elevação do acervo processual de 896 mil processos. O indicador desenvolvido pelo "Justiça em Números" mostra que os processos tramitam em média 4 anos e 3 meses, desde que excluídas as execuções fiscais, cujo tempo é de 6 anos e 9 meses.[185]

A exigência de uma tutela efetiva dos direitos em tempo razoável tem incentivado a desjudicialização no Brasil.[186] O fenômeno global, na verdade, tem impulsionado as partes a selecionar outras formas de solução de conflitos, como a arbitragem, mediação, conciliação e negociação em detrimento da via judicial,[187]

---

[184] CONSELHO NACIONAL DE JUSTIÇA (CNJ). *Barroso recebe líder da Inteligência Artificial do Google e defende uso da IA no Judiciário*. Publicado em 11 jun. 2024. Disponível em: <https://www.cnj.jus.br/barroso-recebe-lider-da-inteligencia-artificial-do-google-e-defende-uso-da-ia-no-judiciario/>. Acesso em 14 jan. 2025.

[185] CONSELHO NACIONAL DE JUSTIÇA (CNJ). *Justiça em números 2024*. Brasília: CNJ, 2024, pp. 15-17. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024.pdf. Acesso em 07 mar. 2025.

[186] THEODORO JÚNIOR, Humberto; ANDRADE, Érico. Novas perspectivas para atuação da tutela executiva no direito brasileiro: autotutela executiva e "desjudicialização" da execução. *Revista de Processo*, vol. 315, p. 109-158, mai. 2021: "(...) não cabe mais, nos dias atuais, falar em monopólio jurisdicional da tutela de direitos, diante da reconhecida "ruína" de tal monopólio, e, além disso, o fato de que tanto no PL 4.257/2019 (relativo à opção da realização da execução fiscal na via administrativa pela própria fazenda pública), como no PL 6.204/2019 (relativo à regulamentação da desjudicialização da execução civil, com a condução do procedimento pelo agente de execução), se insere a regra de que o acesso e o controle jurisdicional continuam sempre abertos, de modo que o devedor ou terceiros podem, a qualquer momento, apresentar impugnações ou mesmo ações judiciais para discutir a execução, o ato de execução ou o direito objeto da tutela executiva extrajudicial. Por isso, como já observava a doutrina italiana, na realidade não se trata de eliminar ou impedir o acesso à jurisdição, mas mudar o momento do controle jurisdicional prévio para o mesmo controle judicial *a posteriori*".

[187] Após o professor Frank Sander verificar a impossibilidade de o Poder Judiciário solucionar, de forma adequada, alguns conflitos familiares, cunhou a necessidade de satisfazer o interesse das partes por outros meios que não a submissão do caso ao juiz. Assim, ele afirmou: "A ideia inicial é examinar as diferentes formas de resolução de conflitos: mediação, arbitragem, negociação e 'med-arb' (combinação de mediação e arbitragem). Procurei observar cada um dos diferentes processos, para ver se poderíamos encontrar algum tipo de taxonomia para aplicar aos conflitos, e que portas seriam adequadas a quais conflitos." (NOGUEIRA, Gustavo Santana; NOGUEIRA, Suzane de Almeida Pimentel. O sistema de múltiplas portas e o acesso à justiça no brasil: perspectivas a partir do novo código de processo civil. *Revista de Processo*, vol. 276, p. 505-522, fev. 2018).

conforme a justiça multiportas adotada pelo próprio Código de Processo Civil.[188] No entanto, parte da doutrina também tem tentado colocar no leque de opções dos cidadãos a possibilidade de efetuarem, eles mesmos, a própria tutela de seus interesses, quando possível.[189] A medida, é claro, encontra resistência.

A doutrina mais conservadora, de forma mais ou menos frequente, tem argumentado que, no ordenamento jurídico atual, o Estado detém o monopólio da força e proíbe o exercício das próprias razões;[190] a autotutela é excepcional e admitida apenas quando expressamente prevista em lei, como, por exemplo, nos casos de urgência causadas pelo inadimplemento de obrigação de fazer ou não

[188] Código de Processo Civil: "Art. 3º Não se excluirá da apreciação jurisdicional ameaça ou lesão a direito. § 1º É permitida a arbitragem, na forma da lei. § 2º O Estado promoverá, sempre que possível, a solução consensual dos conflitos. § 3º A conciliação, a mediação e outros métodos de solução consensual de conflitos deverão ser estimulados por juízes, advogados, defensores públicos e membros do Ministério Público, inclusive no curso do processo judicial."

[189] THEODORO JÚNIOR, Humberto; ANDRADE, Érico. *Novas perspectivas para atuação da tutela executiva no direito brasileiro*..., p. 112: "A autotutela é figura que tradicionalmente tem sido relegada a segundo plano pela doutrina, que se limitava a anotar sua excepcionalidade, na esteira da vedação da atuação privada em tal campo, mas que, na atualidade, vem ganhando novos contornos como importante mecanismo de tutela dos direitos, especialmente quando gerados pela autonomia negocial das partes, instituído pela via contratual, perspectiva em que se destaca a modalidade chamada de 'autotutela executiva', em que, reitere-se, o próprio credor exerce a função executiva-satisfativa da obrigação na via extrajudicial". Em mesmo sentido, mas com base nas lições de Pietro Perlingieri, foi colocado que a autotutela confere às partes o poder de *tutelar* esses interesses diretamente, ao contrário da autonomia privada, que permite aos contratantes *regular* seus próprios interesses (PERLINGIERI, Pietro. *Manuale di diritto civile*. Napoli: ESI, 2005, p. 336. Apud TERRA, Aline de Miranda Valverde Terra. Inafastabilidade da jurisdição e autotutela..., p. 5).

[190] DIDIER Jr., Freddie. *Curso de direito processual civil*. 18. ed. Salvador: JusPodivm, 2016. v. 1, p. 166, afirmar que a autotutela é "solução vedada, como regra, nos ordenamentos jurídicos civilizados. É conduta tipificada como crime: exercício arbitrário das próprias razões (se for um particular) e exercício arbitrário ou abuso de poder (se for o Estado). Como mecanismo de solução de conflitos, entretanto, ainda vige em alguns pontos do ordenamento. Como hipótese excepcional, diz Niceto Alcalá-Zamora y Castillo, a autodefesa é um conceito negativo ou por exclusão". No mesmo sentido, BERMUDES, Sergio. *Introdução ao Processo Civil*. 4. ed. Rio de Janeiro: Forense, 2006, pp. 16-17.

fazer,[191] direito de retenção,[192] desforço possessório,[193] exceção de inadimplemento,[194] compensação,[195] legítima defesa e estado de perigo etc.[196]

Por outro lado, há quem defenda que a autotutela como expressão da autonomia privada, sendo apenas uma das formas de os sujeitos autorregularem seus interesses.[197] É argumentado que, se o indivíduo pode criar normas e estabelecer seus efeitos, deve também poder tutelar os direitos resultantes dessas normas.[198] É preciso avaliar, detidamente, cada um dos argumentos.

Os autores que negam sua aplicação geral argumentam que é preciso reduzi-la para que não exista uma "retração" indevida da jurisdição e, consequentemente, o florescimento de medidas abusivas. No entanto, a manutenção da jurisdição, tal como defendida, pressupõe – por observação da prática – o funcionamento de um Poder Judiciário que tutela interesses legítimos de forma eficiente, o que não se verifica na prática jurídica brasileira, em virtude da morosidade do Judiciário.

---

[191] Código Civil: "Art. 249. Se o fato puder ser executado por terceiro, será livre ao credor mandá-lo executar à custa do devedor, havendo recusa ou mora deste, sem prejuízo da indenização cabível. Parágrafo único. Em caso de urgência, pode o credor, independentemente de autorização judicial, executar ou mandar executar o fato, sendo depois ressarcido. (...) Art. 251. Praticado pelo devedor o ato, a cuja abstenção se obrigara, o credor pode exigir dele que o desfaça, sob pena de se desfazer à sua custa, ressarcindo o culpado perdas e danos. Parágrafo único. Em caso de urgência, poderá o credor desfazer ou mandar desfazer, independentemente de autorização judicial, sem prejuízo do ressarcimento devido."

[192] Um, dentre muitos outros exemplos, Código Civil: "Art. 571. Havendo prazo estipulado à duração do contrato, antes do vencimento não poderá o locador reaver a coisa alugada, senão ressarcindo ao locatário as perdas e danos resultantes, nem o locatário devolvê-la ao locador, senão pagando, proporcionalmente, a multa prevista no contrato. Parágrafo único. O locatário gozará do direito de retenção, enquanto não for ressarcido"

[193] Código Civil: "Art. 1.210. O possuidor tem direito a ser mantido na posse em caso de turbação, restituído no de esbulho, e segurado de violência iminente, se tiver justo receio de ser molestado. § 1º O possuidor turbado, ou esbulhado, poderá manter-se ou restituir-se por sua própria força, contanto que o faça logo; os atos de defesa, ou de desforço, não podem ir além do indispensável à manutenção, ou restituição da posse."

[194] Código Civil: "Art. 476. Nos contratos bilaterais, nenhum dos contratantes, antes de cumprida a sua obrigação, pode exigir o implemento da do outro."

[195] Código Civil: "Art. 368. Se duas pessoas forem ao mesmo tempo credor e devedor uma da outra, as duas obrigações extinguem-se, até onde se compensarem."

[196] Código Civil: "Art. 188. Não constituem atos ilícitos: I - os praticados em legítima defesa ou no exercício regular de um direito reconhecido; II - a deterioração ou destruição da coisa alheia, ou a lesão a pessoa, a fim de remover perigo iminente. Parágrafo único. No caso do inciso II, o ato será legítimo somente quando as circunstâncias o tornarem absolutamente necessário, não excedendo os limites do indispensável para a remoção do perigo."

[197] CABRAL, Antônio do Passo. *Repensando a autotutela*..., p. 32.

[198] SALLES, Raquel Bellini de Oliveira. *Autotutela nas relações contratuais*..., p. 82: "Na dialética do confronto entre velhos e novos perfis da autotutela, pode-se dizer que esta é hoje informada por princípios constitucionais (tais como a dignidade humana, a liberdade, a solidariedade e a livre iniciativa), que, devidamente observados e aplicados, conferem-lhe legitimidade quando a permitem e a expandem, fundamentam a sua inadmissibilidade quando a proíbem ou limitam, e promovem a sua adequação quando a controlam."

Em segundo, o exercício da autotutela exige a presunção de boa-fé por parte dos indivíduos, razão pela qual iniciar a crítica acerca da existência do instituto a partir de seu abuso subverte o próprio sistema normativo.[199] Na verdade, a autotutela, tal como proposta, não seria uma reprodução da vingança privada que foi – ao menos no discurso oficial – erradicada da sociedade com o surgimento da jurisdição moderna. Como qualquer instituto, sua legalidade estaria constrita à ordem constitucional, devendo ser pautada com base nos princípios da razoabilidade e proporcionalidade. Nesse aspecto, a autotutela seria a defesa de interesse legítimo, de forma unilateral e extrajudicial.[200]

Da perspectiva estrutural, portanto, parece que a autotutela tem como fato gerador qualquer fato jurídico (fato jurídico *stricto sensu*, negócio jurídico, ato lícito *stricto sensu* ou ato ilícito) que cause uma lesão real ou potencial na esfera jurídica do indivíduo. Assim, a autotutela é um ato lícito, omissivo ou comissivo, que há de ser tutelado pelo direito, caso tenha finalidade de proteger interesses legítimos ao ordenamento, como ocorre com a legítima defesa, o estado de necessidade e o exercício legítimo do direito no diploma civil (Código Civil, artigo 188, I e II, § único).

---

[199] VENTURI, Thaís Goveia Pascoaloto. *A construção da responsabilidade civil preventiva no direito civil contemporâneo.* (Doutorado). Faculdade de Direito da Universidade Federal do Paraná, Paraná, 2012, p. 236: "Ora, se o que se objetiva assegurar primordialmente é a 'inviolabilidade' e não a 'reparação dos danos causados' à vida, liberdade, igualdade, segurança e propriedade, parece claro que, em relação a tais direitos, antes mesmo de o Estado prestar as mais eficientes formas de tutela imagináveis, abre-se ao próprio titular do direito a autorização, intuitiva até, de que atue, sempre que necessário e urgente, para autotutelar-se contra qualquer ameaça razoável da sua violação. Dessa forma, em que pese todo o preconceito ideológico contra o emprego da autotutela nos sistemas jurídicos atuais, parece certo que o tema merece ser cuidadosamente revisitado, no intuito de se verificar um possível redimensionamento e refundamentação do exercício da autodefesa, sobretudo no que diz respeito à proteção dos direitos fundamentais. A toda evidência, não se deseja defender a volta do emprego da 'vingança privada', da 'força bruta' ou da 'lei do mais forte' para justificar, indevidamente, a autodefesa dos direitos. Muito ao contrário, apenas se suscita a viabilidade de, sem descurar do possível e necessário controle jurisdicional a posteriori, referentemente ao uso arbitrário das próprias razões e ao abuso do direito (a serem viabilizados por via da aplicação de princípios tais como o da razoabilidade, da proporcionalidade e da boa-fé), abrir-se definitivamente o caminho para a aceitação de uma renovada forma de autotutela que, consentânea com os valores e as necessidades da vida social do século XXI, demonstre-se apta a se antecipar ou a complementar a tutela estatal, nem sempre presente, nem sempre acessível, nem sempre célere, nem sempre efetiva."

[200] SALLES, Raquel Bellini de Oliveira. *Autotutela nas relações contratuais*..., p. 82: "Como já explicado anteriormente, a doutrina dedicada ao tema identifica, em geral, três elementos coerentes da noção de autotutela: a defesa de um interesse, a extrajudicialidade e a unilateralidade do comportamento de quem a pratica. Pode-se dizer que os dois primeiros gozam de certa univocidade, reinando alguma controvérsia sobre o terceiro. Com efeito, o problema da unilateralidade do comportamento remete às discussões acerca das possíveis fontes de autotutela e da relação entre esta e a autonomia, bem como da admissibilidade ou não de instrumentos não previstos em lei."

A doutrina também prevê a possibilidade de exercício de autotutela nos casos em haja consentimento das partes e o direito violado seja disponível.[201] Nesse caso, o principal aspecto é que o consentimento do ofendido como excludente de ilicitude é não possuir previsão legal, sendo apenas uma conduta socialmente aceita.

A autonomia negocial permite também que as partes regulem mecanismos de autotutela, mesmo na ausência de previsão legal. A autotutela já é disciplinada para casos em que se envolvem atos ilícitos de qualquer magnitude (de legítima defesa em homicídio até a retirada forçada de invasores da propriedade), sendo necessário apenas o respeito à proporcional e razoável resposta. Não é possível rejeitar, portanto, o projeto "executivo" das partes com base na potencialidade do dano, sob pena de incongruência com o regime jurídico já aplicável e a violação ao princípio da intervenção mínima do judiciário, nos termos do artigo 421, § único, do Código Civil.[202]

A partir dessas premissas, será examinada se a moderação de conteúdo de terceiros pelas empresas de tecnologia (*i.e.*, plataformas digitais) pode ser qualificada como uma sanção contratual pactuada nos termos de uso, que pode ser aplicada pelas plataformas digitais, independentemente da tutela jurisdicional, ganhando contornos dogmáticos de uma autotutela admitida pelo ordenamento jurídico brasileiro, bem como, por consequência, quais as principais consequências da presente qualificação.

2.1.1. Argumentos sobre a moderação de conteúdo pelas redes sociais

A ascensão das redes sociais deu voz a qualquer indivíduo no mundo, permitindo uma forma de autocomunicação de massa.[203] A internet e o surgimento

---

[201] AMARAL, Francisco. Direito civil: introdução. 8ª ed. Rio de Janeiro: Renovar, 2014, p. 582-583: "A respeito da ação ou omissão ilícita do agente, o Código Civil estabelece, no art. 188, hipóteses de especial importância, a legítima defesa e o estado de necessidade como excludentes de ilicitude, isto é, razões que justificam o ato e o tomam lícito. (...) Ainda como excludente de ilicitude, se bem que não-prevista no Código Civil, temos o consentimento do ofendido (*volenti non fit injuria).* Se o prejudicado consente na lesão a seu próprio direito, não há ilicitude no comportamento do agente e o dano não é indenizável. Os direitos atingidos devem ser, porém, disponíveis. Esse princípio releva as lesões que se verificam nas competições esportivas, salvo manifesta intenção de causar dano."

[202] Código Civil: "Art. 421. A liberdade contratual será exercida nos limites da função social do contrato. *Parágrafo único. Nas relações contratuais privadas, prevalecerão o princípio da intervenção mínima* e a excepcionalidade da revisão contratual." (grifou-se)

[203] BALKIN, Jack M., *Free Speech in the Algorithmic Society…*, p. 3.

das redes sociais, embora favoreçam o pluralismo político e as manifestações políticas a nível global, também amplificam o risco de danos causados por discursos abusivos.[204]

A concepção clássica da liberdade de expressão, formulada em um contexto em que a informação era escassa e a participação no debate público requeria altos investimentos financeiros. Nessa perspectiva, predominava a ideia de que a intervenção estatal no discurso seria uma ameaça à autonomia privada e à democracia. A liberdade de expressão era vista como uma liberdade negativa, impondo ao Estado um dever de abstenção.[205]

As redes sociais se tornaram novas praças públicas, criando um espaço de debate *online*, sem controle editorial prévio. O discurso, antes monopólio dos grandes meios de comunicação, foi descentralizado para o espaço digital, democratizando-o, dando voz a grupos minorizados e diversificando o debate público. Em contrapartida, o crescimento das redes sociais e seu uso por bilhões de pessoas no mundo também possibilitou o exercício abusivo da liberdade de expressão, com a disseminação de discursos de ódio, notícias falsas etc.

Kate Klonick afirma que a autorregulação surge no combate à desinformação e à abusividade da liberdade de expressão para assegurar lucros provenientes da publicidade das plataformas.[206] O fundamento da autorregulação,

---

[204] BARROSO, Luís Roberto; BARROSO, Luna van Brussel. Democracia, mídias sociais e liberdade de expressão: ódio, mentiras e a busca da verdade possível. *Direitos Fundamentais & Justiça*, *Belo Horizonte*, ano 17, n. 49, pp. 285-311, jul./dez. 2023: "O mundo vive sob a égide da terceira revolução industrial, também conhecida como a revolução tecnológica ou digital. Algumas de suas principais características são a massificação de computadores pessoais, a universalização dos telefones celulares inteligentes e, acima de tudo, a internet, conectando bilhões de pessoas no planeta. Um dos principais subprodutos da revolução digital e da internet foi o surgimento de plataformas de mídias sociais como o Facebook, Instagram, YouTube, TikTok e aplicativos de mensagens como o WhatsApp e Telegram. Vivemos em um mundo de apps, algoritmos, inteligência artificial e inovação em ritmo acelerado, onde nada parece realmente novo por muito tempo. (...) Antes da internet, a participação no debate público dependia, principalmente, da imprensa profissional, que investigava fatos, seguia padrões da técnica e da ética jornalística e era responsável por danos se publicasse informações falsas, deliberadamente ou por negligência. Havia controle editorial e responsabilidade civil relativamente à qualidade e à veracidade do que era publicado. Isso não significa que fosse um mundo perfeito. O número de meios de comunicação é limitado e nem sempre plural, empresas jornalísticas têm seus próprios interesses e nem todas distinguem com o cuidado necessário fato de opinião. Ainda assim, havia um grau mais refinado de controle sobre o que se tornava público, bem como consequências negativas pela publicação de notícias falsas ou discursos de ódio."

[205] MENDES, Gilmar; BRANCO, Paulo Gustavo Gonet. *Curso de Direito Constitucional*, 9ª ed. São Paulo: Saraiva, 2014, p. 265: "Tratando-se de um típico direito de abstenção do Estado, essa liberdade [de expressão] será exercida, de regra, contra o Poder Público."

[206] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, p. 1.627: "Though corporate responsibility is a noble aim, the primary reason companies

no entanto, não se encontra apenas na lucratividade com publicidade. Afinal, existem indicativos sólidos de que as *fake news* podem, na realidade, trazer significativos retornos financeiros às plataformas digitais.[207] Seja como for, as plataformas que inicialmente eram apenas empresas de tecnologia, passaram a exercer controle sobre o discurso público, tornando-se, para alguns, verdadeiros governantes de espaços digitais.[208]

---

take down obscene and violent material is the threat that allowing such material poses to potential profits based on advertising revenue. Platforms 'sense of the bottom-line benefits of addressing hate speech can be shaped by consumers' — i.e., users' — expectations. If a platform creates a site that matches users' expectations, users will spend more time on the site, and advertising revenue will increase. However, taking down too much content risks losing not only the opportunity for interaction but also the potential trust of users. Likewise, keeping all content up on a site risks making users uncomfortable, leading to a decrease in page views and revenue. Balancing these competing priorities is crucial for platforms aiming to maintain both user trust and profitability." Tradução: "Embora a responsabilidade corporativa seja um objetivo nobre [tornar o mundo mais aberto e conectado], o principal motivo pelo qual as empresas removem materiais obscenos e violentos é a ameaça que esse tipo de conteúdo representa para os lucros potenciais baseados na receita de publicidade. As expectativas dos consumidores — ou seja, dos usuários — podem moldar o 'senso de benefícios financeiros ao lidar com discursos de ódio' das plataformas. Se uma plataforma criar um site que corresponda às expectativas dos usuários, estes passarão mais tempo nele, aumentando a receita publicitária. Contudo, remover muito conteúdo pode resultar na perda não apenas da oportunidade de interação, mas também da confiança potencial dos usuários. Da mesma forma, manter todo o conteúdo disponível em um site pode causar desconforto nos usuários, levando à redução de visualizações de página e de receitas."

[207] Apesar do argumento econômico de que as *fake news* seriam prejudiciais, estudiosos já apontam que "a continuidade da circulação de fake news pode ser relacionada também com o lucro da própria plataforma, como Blotta analisa: 'Existe uma desigualdade no tratamento de diferentes canais: então, se o canal tem muita visualização, eles demoram mais para derrubar e isso provavelmente tem a ver com o retorno financeiro para a plataforma'." (RÁDIO USP. *Alta lucratividade é o que mantém o mercado digital de fake news*. Publicado em 5 dez. 2022. Disponível em: <https://jornal.usp.br/radio-usp/alta-lucratividade-e-o-que-mantem-o-mercado-digital-de-fake-news/>. Acesso em 23 jan. 2025). Em mesmo sentido, SPRING, Mariana. Como usuários do X ganham milhares de dólares espalhando fake news sobre eleição dos EUA. BBC News Brasil, 30 out. 2024. Disponível em: <https://www.bbc.com/portuguese/articles/c937q4p7g09o>. Acesso em 17 jan. 2025.

[208] VIEIRA RAMOS, Carlos Eduardo. *O Direito das Plataformas: Procedimento, legitimidade e constitucionalização na regulação privada da liberdade de expressão na internet*. (Mestrado) Faculdade de Direito da Universidade de São Paulo, São Paulo, 2020, pp. 65 e 67: "(...) para entender como a regulação da liberdade de expressão é transformada no mundo digital, é suficiente compreender como a atuação dos Estados nesse âmbito mudou. E por isso que ele elabora, em síntese, uma taxonomia da atuação estatal, distinguindo-a de acordo com a infraestrutura da liberdade de expressão vigente. Assim, por exemplo, na infraestrutura pré-digital - à qual corresponde a regulação do tipo '*old-school*' -, predominavam atuações diretas do Estado: como ocorreu nos casos *Pentagon Papers* e *Sullivan*, regular a liberdade de expressão se traduzia em estabelecer proibições a que algo fosse divulgado, fazendo-o, por exemplo, via ações judiciais direcionadas contra veículos de imprensa. Na infraestrutura digital - em que criados novos espaços de manifestação, notadamente as plataformas na internet e à qual corresponde a regulação do tipo '*new-school*' – predominam, de modo diferente, atuações indiretas (...) [E]ssa transição entre infraestruturas de liberdade de expressão não necessariamente transforma essas empresas em instrumentos do Estado, mas frequentemente as coloca na posição de decidir, de forma independente, controvérsias quanto à liberdade de expressão. (...) A partir de uma reconstrução do mecanismo decisório criado para atender a essa demanda – a moderação de conteúdo – que os impactos do mundo digital na regulação da liberdade de expressão não se traduziram apenas em

Agora, a questão principal é: qual é o fundamento jurídico da moderação de conteúdo ou da remoção de contas nas redes sociais?

A título de exemplo, o *Instagram* possui poderes significativos na moderação de conteúdos de terceiros em sua plataforma, conforme descrito nos termos de uso.[209] A empresa pode remover qualquer conteúdo ou informação que julgue violar seus termos, políticas ou leis aplicáveis, visando a proteger a comunidade e garantir o funcionamento seguro e eficiente do serviço.[210] A remoção de conteúdo pode ocorrer se a publicação for considerada abusiva, enganosa, ilegal ou prejudicial de qualquer forma, tendo poderes também para desativar ou encerrar contas que desrespeitem repetidamente seus termos ou que criem riscos legais para a empresa.

No referido negócio, são destacadas várias circunstâncias específicas em que a empresa pode agir contra conteúdos e contas, como infrações às suas "Normas da Comunidade", que incluem regras contra comportamentos abusivos, *bullying*, assédio, incitação ao ódio, fraudes, violação de direitos de propriedade intelectual, disseminação de informações privadas ou confidenciais sem permissão, comportamento de *spam*, criação de contas falsas, criação de contas através de meios automatizados ou para fins de venda e compra de contas, prática de *spam*, incluindo o envio de mensagens não solicitadas em massa e assim por diante.[211]

---

mudanças nas técnicas utilizadas pelo Estado. Pelo contrário: eles também significaram a criação de novos espaços de decisão, materializados em nichos de poder privados que, localizados em companhias como o Facebook e o Google, regulam a liberdade de expressão ao determinar, com independência, aquilo que as pessoas podem ou não podem dizer em importantes – se não os mais importantes – espaços de manifestação atual: as plataformas da *internet*."

[209] INSTAGRAM. Termos de utilização. Publicado em 26 jul. 2022. Disponível em: <https://help.instagram.com/581066165581870/?locale=pt_PT&hl=pt>. Acesso em 13 jan. 2025.

[210] A mudança de política anunciada pela Meta por Mark Zuckerberg pode alterar significativamente esses poderes. Ver, para mais detalhes, o subcapítulo 1.7 *supra*.

[211] Na tentativa de contornar a moderação de conteúdo, o ex-presidente Jair Bolsonaro editou a Medida Provisória 1.068/2021 para impedir o cancelamento de perfis ou a retirada de conteúdo postado por terceiros, restringindo a atuação das empresas aos temas de nudez, pedofilia ou terrorismo. A situação é de toda similar ao caso *Moody, Attorney General of Florida, et al. v. NetChoice, LLC, DBA NetChoice, et al.*, julgado pela Suprema Corte dos Estados Unidos. Aqui, a Corte examinou a constitucionalidade de uma lei da Flórida que restringia as práticas de moderação de conteúdo de grandes plataformas digitais. A lei buscava impedir que essas plataformas banissem ou diminuíssem a visibilidade de conteúdos com base em critérios políticos, sob o argumento de proteger a liberdade de expressão dos usuários. Empresas como NetChoice argumentaram que a lei violava a Primeira Emenda ao limitar a autonomia das plataformas de regular conteúdos em seus espaços digitais. A Corte reconheceu que as plataformas digitais, embora sejam espaços de comunicação amplamente utilizados, possuem direitos de liberdade de expressão. Assim, decisões sobre moderação de conteúdo fazem parte de sua autonomia editorial. Contudo, a decisão também indicou que estados podem, em determinados contextos, impor regulamentações para evitar discriminação arbitrária e proteger interesses públicos, desde que respeitem os limites

Existe, portanto, um negócio jurídico que regula de forma específica a relação entre os usuários e as empresas de tecnologia (plataformas digitais). A análise dos Termos de Uso, ao que tudo indica, aponta não apenas quais serviços são prestados pela plataforma digital e a extensão dos direitos detidos pelos indivíduos, como também os mecanismos de execução forçada em caso de inadimplemento.

Com fundamento nas regras contratuais, previstas em termos de uso previamente aceitos pelos usuários, as plataformas impõem suspensões e/ou exclusões de contas e/ou publicações de usuários das redes sociais. Trata-se de sanções contratuais que são aplicadas por meio de sistemas informatizados próprios, sem a necessidade de as plataformas recorrerem à tutela jurisdicional. A distinção desse modelo decorre justamente da ausência da intervenção de um terceiro para a satisfação da pretensão das plataformas digitais.

Por exemplo, a multa por descumprimento contratual é uma sanção prevista em um negócio jurídico, mas exige o envolvimento do Estado-juiz para que seja executada pela parte credora da obrigação inadimplida. Assim, o terceiro (Estado-juiz) intervém para satisfazer a pretensão da parte credora. No caso das plataformas, as redes sociais possuem todos os meios técnicos para realizar a moderação de conteúdo a partir da aplicação dos termos de uso. Elas não precisam da intervenção do Estado-juiz para que seja satisfeita sua pretensão de assegurar, por exemplo, que o ambiente da plataforma esteja de acordo com as suas regras. As empresas de tecnologia exercem, de forma unilateral, a defesa dos interesses que elencou como legítimos em seus termos de uso e, por via extrajudicial, busca defendê-los com a aplicação de sanções contratuais. Logo, é razoável entender que as plataformas exercem a autossatisfação de seu direito à liberdade editorial.

Esse modelo de gestão interna é vital para a rápida resolução de problemas, adaptando-se às mudanças constantes e dinâmicas do ambiente digital. A empresa se baseia em seus termos de uso, que os usuários concordam ao utilizar o serviço, para justificar suas ações e proteger tanto os indivíduos quanto a própria integridade da plataforma. Essa prática se tornou fundamental ao Estado brasileiro, como ficou demonstrado pelo acordo celebrado entre o Tribunal Superior Eleitoral

---

constitucionais (SUPREME COURT OF THE UNITED STATES. *Moody, Attorney General of Florida, et al. v. NetChoice, LLC, DBA NetChoice, et al.* Disponível em: <https://www.supremecourt.gov>. Acesso em 13 jan. 2025).

e as empresas Meta, em 2020, que efetuaram controle prévio da liberdade de expressão para impedir danos ao sistema eleitoral.[212]

Isso abre margem, portanto, ao debate que ora se propõe neste trabalho pela qualificação dogmática dessa atividade como uma autotutela, exercida pelas plataformas digitais, para manter o debate público e atividade econômica virtual minimamente ordenados.

Afinal, a autossatisfação do direito à liberdade editorial pelas plataformas digitais, por meio da aplicação forçada das sanções contratuais previstas nos termos de uso, sem que seja necessário recorrer à tutela jurisdicional, poderia, portanto, ser considerada uma espécie de autotutela?

Ante a peculiaridade das redes sociais, as plataformas teriam o dever de exercer essa autotutela? Quais parâmetros normativos devem guiar as condutas das redes sociais em relação à liberdade de expressão? Podem essas empresas ser responsabilizadas por remover conteúdo ou excluir contas, mesmo que não seja ilegal, ao aplicar seus termos de uso?[213]

A dicotomia entre a restrição da autotutela tradicional e sua prática moderna sugere uma necessidade de reavaliação dos fundamentos jurídicos que governam esse instituto. A doutrina tem se esforçado para incorporar a autotutela como parte integrante da ordem geral da tutela dos direitos, reconhecendo sua importância na proteção de interesses legítimos em situações em que a intervenção judicial é impraticável ou ineficaz. Essa abordagem pragmática busca equilibrar a

---

[212] Tribunal Superior Eleitoral (TSE). TSE assina parceria com Facebook Brasil e WhatsApp Inc. para combate à desinformação nas Eleições 2020. Publicado em 30 set. 2020. Disponível em: <https://www.tse.jus.br/comunicacao/noticias/2020/Setembro/tse-assina-parceria-com-facebook-brasil-e-whatsapp-inc-para-combate-a-desinformacao-nas-eleicoes-2020>. Acesso em 14 jan. 2024.

[213] O MCI prevê que as empresas devem ser comprometidas em preservar a liberdade de expressão, só sendo responsabilizadas pelos conteúdos de terceiros quando não removerem o conteúdo ilícito no prazo determinado pelo juízo. A moderação de conteúdo, assim, não seria, em tese, exigível sem a judicialização prévia, demonstrando, ainda mais, a subversão normativa e lógica do art. 19 do MCI. O dispositivo não só é objeto de contundentes críticas doutrinárias, mas sobretudo de questionamentos quanto à sua constitucionalidade perante o STF, por conter, no artigo 19, norma de limitação da responsabilidade civil dos provedores de internet por conteúdo gerado por terceiros, com a finalidade expressa no próprio dispositivo de "assegurar a liberdade de expressão e impedir a censura". Para mais informações, remeta-se a PEREIRA DE LIMA, Cíntia Rosa; FRANCO DE MORAES, Emanuele Pezati; PEROLI, Kelvin. *O necessário diálogo entre o Marco Civil da Internet e a Lei Geral de Proteção de Dados para a coerência do sistema de responsabilidade civil diante das Novas Tecnologias*, In Responsabilidade civil e novas tecnologias; coordenado por Guilherme Magalhães Martins. Nelson Rosenvald – Indaiatuba, SP. Editora Foco, 2020, pp. 146; e SCHREIBER, Anderson. *Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro*., In: LUCCA, Newton de; SIMÃO FILHO, Adalberto; LIMA, Cíntia Rosa Pereira. Direito e Internet III: Marco Civil da Internet, Lei nº 12.965/2014, Tomo II. São Paulo: Quartier Latin, 2015, pp. 277-305.

necessidade de intervenção mínima do judiciário com a realidade das relações sociais e econômicas instantâneas e complexas da era digital.

Portanto, é fundamental que o direito evolua para acomodar as novas realidades impostas pela tecnologia e pela globalização, reconhecendo uma possível autotutela como um mecanismo válido e necessário de resolução de conflitos.[214] As empresas de tecnologia, ao exercerem a autossatisfação de seu direito à liberdade editorial, com a aplicação forçada de seus termos de uso, exemplificam o que pode ser uma adaptação moderna desse instituto, desempenhando um papel crucial na manutenção da ordem e na proteção dos direitos no espaço digital. A discussão contínua sobre os limites e a legitimidade da autotutela é essencial para garantir que ela seja utilizada de maneira justa e proporcional, protegendo os interesses individuais e coletivos em um mundo cada vez mais interconectado.[215]

2.2. Moderação de conteúdo nas redes sociais com base nos termos de uso

As redes sociais podem e devem realizar a moderação de conteúdo. Essa inafastável conclusão independe se as plataformas desempenham de forma adequada ou inadequada a curadoria do conteúdo *online*. Partindo dessa premissa, é preciso que seja reconhecido os direitos editoriais das plataformas digitais, que são uma extensão dos direitos à liberdade de expressão.[216]

---

[214] CABRAL, Antônio do Passo. *Repensando a autotutela*..., pp. 33-34: "Nesse cenário, outro fator que tem contribuído para a reabilitação da autotutela é o uso da tecnologia para a formação das relações jurídicas contratuais, que estrutura mecanismos de autotutela não associados a formas violentas de solução de disputas. Em especial, cabe destacar os *smart contracts* ou contratos inteligentes, cujos exemplos práticos já se tornam realidade para muitos brasileiros."

[215] *Ibidem*, pp. 37-38: "A autotutela implica o exercício concreto e unilateral de poder, e, portanto, é relevante que o sistema jurídico estabeleça restrições para essas práticas. É preciso, por conseguinte, investigar os limites para o exercício da autotutela, balizamentos que devem iluminar o controle judicial desses atos. Ressalto que, como forma de solução de conflitos, que pode levar à supressão do patrimônio, a autotutela precisa ser processualizada, e o conjunto de atos que compõem este iter deve atender às exigências do que se poderia chamar de um devido processo da autotutela, parâmetros que devem ser seguidos para garantir que a autotutela não seja abusiva. Especificamente em relação às exigências materiais no exercício da autotutela, entendo que é possível extraí-las do sistema jurídico, em especial de algumas previsões já tipificadas em lei. Isso porque o legislador, ao disciplinar algumas modalidades de autotutela, fornece ao intérprete pistas de quais são os limites para que seu desempenho concreto seja lícito."

[216] BALKIN, Jack M. *How to Regulate (and Not Regulate) Social Media*..., p. 90: "Social media platforms must engage in content moderation. They may do it badly or well, but they will have to do it nevertheless. Accordingly, governments should respect social media's role as curators and editors of public discourse. Respecting that role means that social media should have editorial rights,

A liberdade de expressão das redes sociais, por meio da liberdade editorial, deve ser reconhecida e encarada como um limite regulatório a ser considerado, garantindo-lhe a escolha e a possibilidade de moderar conteúdos, a fim de que seus usuários tenham acesso àquilo que as redes sociais entendem que eles querem ver em suas plataformas.[217]

Até recentemente, a maioria dos países havia consolidado a premissa de que as plataformas digitais não teriam capacidade real de monitorar, diariamente, a vasta gama de discursos publicados nas redes sociais e identificar de forma precisa a ilegalidade do conteúdo para moderá-lo. Essa premissa constava, inclusive, das regulações norte-americanas e europeia, DMCA and *EU eCommerce Directive*, que expressamente limitavam as obrigações de monitoramento, responsabilizando-as apenas caso comprovado o prévio conhecimento a respeito do conteúdo ilegal, o que normalmente ocorria por meio de notificação de terceiros.[218]

A percepção inicial quanto à impossibilidade técnica das plataformas de realizar o monitoramento dos conteúdos postados nas redes sociais está presente no contexto brasileiro. Conforme tratado no subcapítulo 1.5 *supra*, tanto a jurisprudência do STJ que havia se formado antes da vigência do Marco Civil da Internet – a responsabilidade civil das plataformas dependia do prévio conhecimento (*Notice and Takedown*) –, quanto as normas do próprio diploma – que condicionam a responsabilidade ao recebimento da notificação judicial que determine a retirada de conteúdo (*Judicial Notice and Takedown*) – não estipulam qualquer obrigação de monitoramento prévio ou de moderação de conteúdo.

Ao longo dos anos, as plataformas digitais passaram a desenvolver voluntariamente sofisticados mecanismos de filtragem tecnológica de conteúdo, que simplesmente alteraram as expectativas dos legisladores, embora o modo de funcionamento dessas ferramentas ainda seja pouco compreendido. A partir dessa mudança de perspectiva, novas leis passaram a efetivamente exigir a filtragem ou o monitoramento prévio do conteúdo, sendo um dos exemplos dessa nova legislação,

---

which are a subset of free speech rights." Tradução: "As plataformas de rede social devem moderar conteúdo. Elas podem fazer isso bem ou mal, mas terão que fazer isso mesmo assim. Assim, os governos devem respeitar o papel da rede sociais como curadores e editores do discurso público. Respeitar esse papel significa que a rede social deve ter direitos editoriais, que são um subconjunto dos direitos de liberdade de expressão."

[217] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., pp. 43-104, p. 59.

[218] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* …, p. 9.

a *EU Copyright Directive*, que dispõe sobre direitos autorais e direitos conexos ao mercado único digital na União Europeia.[219]

Se, por um lado, a regulação, em boa parte do mundo, inclusive nos Estados Unidos (CDA) e no Brasil (MCI), não obriga as plataformas a realizarem o monitoramento prévio de conteúdo, garantindo-as imunidade com o intuito de promover o aperfeiçoamento de seu sistema de autorregulação, por outro, é fato que as plataformas digitais promovem ativamente a moderação do conteúdo *online*.

Além do regime de autorregulação, Kate Klonick elenca duas razões para explicar por que as plataformas criam regras e um sistema de curadoria *online* para moderar conteúdo obscenos, violentos ou de discurso de ódio: *(i)* um senso de responsabilidade corporativa, alinhados com valores da liberdade de expressão e da democracia;[220] e, sobretudo, *(ii)* para evitar que esses materiais ameacem os lucros potenciais com base na receita de publicidade.[221]

As plataformas têm um "legítimo interesse de evitar que as suas comunidades sejam dominadas por conteúdos como esses", porque "a grande maioria dos usuários, que não quer ser exposta a esse tipo de conteúdo, seria afugentada das plataformas", causando relevante impacto financeiro pela perda de receita publicitária. Por isso, a moderação de conteúdo é "uma atividade intrínseca ao serviço oferecido por essas empresas".[222]

---

[219] *Idem*.

[220] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … pp. 1.625-1.626: "Some platforms choose to moderate content that is obscene, violent, or hate speech out of a sense of corporate responsibility". Tradução: "Algumas plataformas optam por moderar o conteúdo obsceno, violento ou de discurso de ódio por um senso de responsabilidade corporativa".

[221] *Ibidem*, p. 1.627: "Though corporate responsibility is a noble aim, the primary reason companies take down obscene and violent material is the threat that allowing such material poses to potential profits based in advertising revenue. Platforms 'sense of the bottom-line benefits of addressing hate speech can be shaped by consumers' — i.e., users' — expectations.' If a platform creates a site that matches users' expectations, users will spend more time on the site and advertising revenue will increase. Take down too much content and you lose not only the opportunity for interaction, but also the potential trust of users. Likewise, keeping up all content on a site risks making users uncomfortable and losing page views and revenue." Tradução: "Embora a responsabilidade corporativa seja um objetivo nobre, a principal razão pela qual as empresas retiram material obsceno e violento do ar é a ameaça que a permissão desse material representa para os possíveis lucros baseados na receita de publicidade. O 'senso das plataformas sobre os benefícios finais de lidar com o discurso de ódio pode ser moldado pelas expectativas dos consumidores, ou seja, dos usuários'. Se uma plataforma criar um site que corresponda às expectativas dos usuários, eles passarão mais tempo no site e a receita de publicidade aumentará. Retirar muito conteúdo e você perderá não apenas a oportunidade de interação, mas também a possível confiança dos usuários. Da mesma forma, manter todo o conteúdo de um site corre o risco de deixar os usuários desconfortáveis e perder visualizações de página e receita."

[222] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo...*, p. 221.

A ideia central deste capítulo não perpassa apenas pela existência de um dever ou de um direito das plataformas digitais de moderar o conteúdo *online*, mas sobretudo de que as redes sociais já o fazem – e devem adotar procedimentos confiáveis e transparentes no âmbito dos processos decisórios sobre moderação de conteúdo, para garantir a legitimidade da moderação que já é feita.

As redes sociais promovem diariamente o monitoramento *online* por meio de diversos mecanismos e procedimentos, que envolvem o controle prévio (moderação *ex ante*) e posterior à postagem do conteúdo (moderação *ex post*). A moderação de conteúdo pode ocorrer de forma reativa, na qual os moderadores avaliam passivamente o conteúdo após usuários denunciarem determinado conteúdo publicado (*flagging*), ou proativamente, quando os moderadores procuram ativamente por conteúdo para remoção. Essa moderação pode, ainda, ser feita por inteligência artificial ou revisores humanos.

As redes sociais desempenham também a curadoria do conteúdo *online* por meio de mecanismos de filtragem algorítmica para modular a experiência individual do usuário. Os algoritmos customizam o conteúdo exibido, tornando visíveis ou invisíveis determinados conteúdos, de acordo com as preferências e os hábitos dos usuários. Em termos práticos, essa tecnologia pode garantir visibilidade e promover determinados conteúdos ou, por outro lado, torná-los invisíveis ao ponto de equivaler à proibição de acesso.

Embora a filtragem algorítmica para modular a experiência do usuário suscite questões a respeito de sua neutralidade,[223] aliado ao desconhecimento da literatura sobre seu funcionamento, por permear a proteção a segredos de negócios e operar por meio de um mecanismo de aprendizagem de máquina (*machine learning*),[224] o uso de algoritmos para customização dos conteúdos não está inserido

[223] NITRINI, Rodrigo Vidal. *Liberdade de Expressão nas Redes Sociais: o problema jurídico da remoção de conteúdo pelas plataformas...*, p. 79: "*As grandes redes sociais não são ambientes neutros*, onde prevaleceria uma irrestrita possibilidade publicação por indivíduos. Se é verdade que há uma liberdade sem precedentes de publicação praticamente imediata em seus ambientes, ela convive com *diversos e relevante mecanismos de filtragem que operam entre os campos do permitido/proibido e do visível/invisível*. A derrubada de conteúdo pode ser uma intervenção mais extrema, mas o próprio funcionamento regular das redes sociais pressupõe que a plataforma escolha o nível de visibilidade/invisibilidade de veiculação das postagens, conforme sua *curadoria algorítmica*."

[224] Sobre a filtragem algorítmica, Rodrigo Nitrini adverte que "é também a tecnologia mais opaca e blindada do debate público, em parte porque muitas vezes os algoritmos são protegidos como segredos de negócios, em parte também porque, mesmo quando esse não é o caso, sua operação é tão complexa a ponto de dificultar sua compreensão por pessoas sem formação técnica especializada.

nas operações realizadas pelas plataformas digitais com base nas políticas de moderação de conteúdo dispostas nos seus Termos de Uso. Por essa razão, o enfrentamento da problemática também aponta para soluções regulatórias que se distanciam do escopo da análise deste estudo, em especial a desmonetização de conteúdo e a regulação do conteúdo publicitário.

De volta à questão central deste capítulo, não se pode distanciar da complexa realidade que envolve a moderação de conteúdo. É fundamental que qualquer solução normativa que vise a regulamentar a moderação de conteúdo feita pelas redes sociais considere a complexidade desse contexto tecnológico, para compreensão dos riscos inerentes à regulação e das necessidades que se busca enfrentar. Assim, será possível construir caminhos normativos que promovam os direitos da liberdade de expressão essenciais à preservação da cultura democrática.

Para desenvolver as operações de moderação de conteúdo nas redes sociais, passa-se a apresentar as estruturas de controle do discurso público *online* nas redes sociais com base na moderação de conteúdo realizada pelas três maiores plataformas transnacionais, *Facebook*, *Youtube* e *X* (antigo *Twitter*).

2.2.1. A moderação de conteúdo das plataformas digitais do *Facebook*, *Youtube* e *X*: a evolução das diretrizes gerais para o sistema de regras

As três maiores plataformas transacionais de redes sociais, *Facebook*, *Youtube* e *X*, desempenham um papel central na moderação do conteúdo compartilhado por bilhões de usuários em todo o mundo.

O *Facebook,* controlado pela *Meta Platform Inc.* ("Meta"), é a maior rede social da atualidade, tendo atingido 3,07 bilhões de usuários mensais ativos no último trimestre de 2023.[225] O *Youtube*, controlado pelo *Google LLC* ("Google"), segue como a segunda rede social mais acessada no mundo, com quase 2 bilhões

Algoritmos, por exemplo, não se limitam a aplicar instruções previa e expressamente definidas em sua programação, pois podem operar por meio de 'aprendizagem automática' ('machine learning'), que evoluem a partir de seu próprio uso ou da alimentação de dados, incluindo a interação entre algoritmos distintos". (*Ibidem*, pp. 71-72).

[225] DIXON, Stacy Jo. Number of monthly active Facebook users worldwide as of 4th quarter 2023. *STATISTA*, publicado em 21 mai. 2024. Disponível em: <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>. Acesso em 9 dez. 2024.

de usuários mensais ativos em maio de 2023.[226] Por sua vez, o X (antigo Twitter), detido pela X Corp., contabilizou 540 milhões de usuários mensais em julho de 2023.[227] Ao longo dos últimos anos, essas plataformas se consolidaram como dominantes no compartilhamento global de conteúdo *online*.[228]

A escolha pela análise específica dessas plataformas transnacionais para expor *como* é feita a moderação de conteúdo nas redes sociais não é meramente quantitativa. Essa opção está baseada, sobretudo, na relevância e na abrangência da moderação de conteúdo que elas exercem. A título de exemplo, no último semestre de 2020, o *Facebook* moderou mais de 105 milhões de conteúdos publicados em sua plataforma, uma média de 1,1 milhão por dia. No mesmo período, o *Youtube*, que recebe 500 horas de novos vídeos por minutos, removeu mais de 9,3 milhões de vídeos. Já o X analisou mais de 12,4 milhões de contas por violações às suas regras no primeiro semestre de 2020.[229] Esses exemplos revelam a complexidade dos desafios que essas plataformas enfrentam na moderação de conteúdo, diante da escala massiva de informações compartilhadas nas redes sociais.

Entretanto, essa realidade não se estabeleceu sem um percurso significativo. Historicamente, a moderação de conteúdo nessas plataformas digitais teve início por meio do estabelecimento de diretrizes simplificadas e pouco claras e que, por muitas vezes, levavam em conta o julgamento subjetivo de um número reduzido de moderadores. Até outubro de 2006, quando foi adquirido pelo *Google*, o *Youtube* não havia implementado suas políticas de moderação de conteúdo, quando, então, especialistas foram contratados para estabelecer diretrizes de moderação, que tinham como base a proteção da liberdade de expressão e vedação à censura.[230]

---

[226] FORBES. *Brasil é o terceiro país com mais usuários do YouTube em 2023*. *Forbes*, Publicado em 10 mai. 2023. Disponível em: <https://forbes.com.br/forbes-tech/2023/05/brasil-e-o-terceiro-pais-com-mais-usuarios-do-youtube-em-2023/>. Acesso em 9 dez. 2024: "Criado em 2005 e disponível em mais de 100 países, o YouTube é até hoje o segundo site mais acessado do mundo, ficando atrás somente do Google, e possui cerca de 2 bilhões de usuários ativos mensalmente..."

[227] REUTERS. Musk diz que plataforma X atingiu novo recorde de usuários mensais. *Reuters*, publicado em 29 jul. 2023. Disponível em: <https://www.infomoney.com.br/negocios/musk-diz-que-plataforma-x-atingiu-novo-recorde-de-usuarios-mensais/>. Acesso em 9 dez. 2024.

[228] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.603.

[229] DOUEK, Evelyn. *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability*. Columbia Law Review, v. 121, n. 3, 2021, p. 759-834, p. 791. Disponível em: <https://ssrn.com/abstract=3679607>. Acesso em 9 dez. 2024.

[230] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … pp. 1.619-1.620: Eventos que ocorreram em 2016 e nos anos seguintes – como a divulgação de vídeos da execução de Saddam Hussein, em dezembro de 2006, de um homem sendo brutalmente espancado pela polícia no Egito (postado por ativista egípcio para denunciar as

De forma similar, o *Facebook* sequer havia publicado, até novembro de 2009, os seus "Padrões da Comunidade" ("*Community Standards*"). As decisões de moderação de conteúdo eram, até então, baseadas em uma restrita e pouca clara listagem de diretrizes internas. Dotada de caráter primordialmente subjetivo, essas diretrizes internas consistiam basicamente na orientação aos moderadores de que o conteúdo postado deveria ser removido "se algo faz você se sentir mal".[231] A partir de 2009, o *Facebook* formou uma equipe especializada em moderação de conteúdo, que começou uma nova fase na curadora do conteúdo *online*, estruturando e publicando uma política formal de regras, os seus "Padrões da Comunidade"[232].

A aplicação de diretrizes gerais evoluiu para um regime de regras, que refletem comandos mais específicos e precisos sobre como os moderadores devem analisar e moderar o conteúdo *online*.[233] Essa transição foi feita pelo *YouTube* e pelo *Facebook*, em virtude dos seguintes fatores: *(i)* o aumento rápido do número de usuários e do volume de conteúdo postado, *(ii)* a globalização e a diversidade da comunidade *online*, bem como *(iii)* a crescente dependência de equipes de moderadores humanos provenientes de diversos contextos culturais.[234]

Portanto, à medida que o nível de complexidade das comunicações nas redes sociais aumentou significativamente com o passar dos anos, tornou-se evidente a necessidade de criar um conjunto mais robusto de regras para lidar com o crescente volume e a diversidade dos conteúdos publicados *online*.

Ao contrário do *YouTube* e do *Facebook*, que implementaram processos ativos de moderação de conteúdo, o *X* (antigo *Twitter*) não desenvolveu

---

violações de direitos humanos), em 2017, e de um manifestante do Movimento Verde iraniano que foi baleado durante um protesto, em junho de 2009 –, foram fundamentais para a definição de novas políticas e diretrizes internas no *Youtube*, em especial sobre a publicação de conteúdo violento no *Youtube*.

[231] *Ibidem*, p. 1631.

[232] *Ibidem*, p. 1620.

[233] Kate Klonick explica que a escolha entre adotar um sistema de regras ou padrões traz implicações significativas, impactando a execução das normas e a justiça percebida nas decisões de moderação. De acordo com a autora, as diretrizes são mais geral e permitem interpretações e aplicações variadas, sendo um exemplo delas "não dirigir muito rápido", enquanto as regras são comandos específicos e exatos que os moderadores devem seguir, como não dirigir acima do limite de velocidade de 65 milhas por hora. A autora complementa que a adoção de um sistema ou outro possui vantagens e desvantagens. As diretrizes, embora mais vagas e sujeitas ao arbítrio do moderador, são mais adaptáveis à mudança de contexto, característica que é fundamental para curadoria *online*. As regras, por outro lado, podem ser mais fáceis de aplicar pelos moderadores, mas resultar em resultado injustos, em virtude da baixa discricionariedade e das limitações decorrentes de eventuais lacunas (*Ibidem*, pp. 1.631-1.632).

[234] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1635.

inicialmente um sistema interno de regras para remover ou revisar postagens *online*. Ao revés, o X adotou uma política inicial de não moderar o conteúdo dos usuários, em prol de um padrão de liberdade de expressão que se baseou tanto na não intervenção em conteúdo dos usuários quanto na proteção desses conteúdos.[235]

De fato, comparativamente ao *Facebook* e ao *Youtube*, o *X*, então *Twitter*, foi considerado por muitos anos "uma plataforma mais propícia a discursos agressivos e de ódio", em virtude do próprio *design* de suas regras de conversação, que não permitia que um usuário removesse de suas postagens reações de terceiros, que se tornavam públicas a todos que visualizassem o conteúdo postado.[236]

A mudança veio apenas a partir de 2015, quando o *X* se comprometeu a se engajar na curadoria do conteúdo *online* e, nos anos seguintes, implementou políticas e ferramentas para facilitar aos usuários os usuários filtrem e ocultem o conteúdo que não querem visualizar.[237] Dentre essas ferramentas, a empresa disponibilizou, no ano de 2019, uma opção de bloqueio ou ocultação de comentários pelo autor da postagem original, com a intenção de incentivar a manutenção de um ambiente de discurso mais saudável e civilizado.[238]

A moderação de conteúdo nas três maiores plataformas transnacionais de rede sociais, *Facebook*, *Youtube* e *X*, é um campo dinâmico que envolve a constante adaptação a desafios emergentes. À medida que essas plataformas continuam a evoluir no campo da governança privada do discurso público, o equilíbrio entre a promoção dos valores da liberdade de expressão e a necessidade de proteger o bem-estar da comunidade digital permanecerá na centralidade do debate sobre a regulação das redes sociais.

Para melhor compreensão desse desafio, o capítulo seguinte busca retratar as formas de moderação de conteúdo com enfoque na aplicação das regras elaboradas pelas plataformas *Facebook, Youtube* e *X*, assim como nos problemas e dilemas associados à implementação dos procedimentos técnicos de moderação a seguir desenvolvidos. A análise desses procedimentos revelará que existem hoje diversas formas de moderação de conteúdo que vão além das formas tradicionais –

---

[235] *Ibidem*, p. 1621.
[236] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., pp. 48-49.
[237] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … pp. 1627.
[238] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 49.

representados na "velha escola" da regulação do discurso –, que, conforme explica Jack M. Balkin, estavam relacionadas diretamente aos palestrantes e editores de conteúdo, sujeitos à aplicação de uma série de sanções, desde multas civis a prisões, para censurá-los.[239]

2.2.1.1. Controle automatizado de conteúdo prévio à publicação

O controle automatizado de conteúdo prévio à publicação é um mecanismo de moderação *ex ante* que consiste uma exceção à regra de que os usuários podem postar seus conteúdos nas plataformas digitais sem avaliação prévia e, até mesmo, remete à ideia de censura prévia, uma vez que boa parte da moderação ocorre após a publicação.[240]

A avaliação prévia é feita no intervalo entre o *upload* de um vídeo e sua publicação em plataformas como o *Facebook* e o *Youtube*. Quando um usuário carrega um vídeo no *Facebook*, por exemplo, recebe uma notificação indicando que o conteúdo está sendo processado. É nesse momento que tem início a moderação automatizada, por meio da triagem algorítmica, para avaliar conteúdos antes da publicação, sem a revisão humana.[241]

Esse procedimento de moderação é utilizado de forma mais habitual no combate à pornografia infantil, ao qual as plataformas digitais são obrigadas a fazer. Kate Blonick explica que a *Section* 230 do CDA expressamente estabelece que os provedores de internet não têm imunidade para o descumprimento de leis federais

---

[239] BALKIN, Jack M., *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation…*, pp. 27-28: "Old school speech regulation primarily aims at speakers and publishers of content. It uses traditional methods of enforcement, including civil and criminal fines, injunctions, imprisonment, and in some countries, violence or the threat of violence to deter and censor speakers and publishers. New school speech regulation, by contrast, is not aimed at speakers or publishers; it is aimed at digital infrastructure. What is the digital infrastructure? It includes the Internet backbone, cloud services, the international domain name system ("DNS"), Internet service providers, web hosting services, social media platforms, and search engines." Tradução: "A regulamentação do discurso da velha escola visa principalmente aos oradores e editores de conteúdo. Ela usa métodos tradicionais de aplicação, incluindo multas civis e criminais, liminares, prisão e, em alguns países, violência ou ameaça de violência para deter e censurar palestrantes e editores. A regulamentação do discurso da nova escola, por outro lado, não é direcionada a palestrantes ou editores; ela é direcionada à infraestrutura digital. O que é a infraestrutura digital? Ela inclui o backbone da Internet, serviços em nuvem, o sistema internacional de nomes de domínio ("DNS"), provedores de serviços de Internet, serviços de hospedagem na Web, plataformas de mídia social e mecanismos de busca".

[240] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas...*, p. 50.

[241] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.636.

criminais nos Estados Unidos da America, como proteção à propriedade intelectual ou privacidade das comunicações. Isso significa que, por força de leis federais estadunidenses, as plataformas digitais devem combater a pornografia infantil.[242]

Como as três maiores plataformas transnacionais objeto deste estudo, *Facebook*, *Youtube* e *X*, estão situadas no Estados Unidos da América, as regras de moderação dessas plataformas estão intrinsecamente ligadas à legislação e ao entendimento da jurisprudência estadunidenses. A título de exemplo, o *Facebook* expressamente proíbe, em suas "Padrões de Comunidade", a publicação de qualquer conteúdo relacionado à exploração sexual infantil.[243]

Para o cumprimento desse dever, as plataformas digitais, como *Facebook*, utilizam algoritmos de reconhecimento de imagem, como o *PhotoDNA*, que desempenham um papel central no combate à pornografia infantil, por meio de mecanismo de "impressão digital" (*hash*).[244] O algoritmo identifica e bloqueia

---

[242] *Idem*: "It is important to remember that § 230 expressly states that no internet entity has immunity from federal criminal law, intellectual property law, or communications privacy law. 47 U.S.C. § 230(e) (2012). This means that every internet service provider, search engine, social networking platform, and website is subject to thousands of laws, including child pornography laws, obscenity laws, stalking laws, and copyright laws". Tradução: "É importante lembrar que o § 230 estabelece expressamente que nenhum provedor de Internet tem imunidade em relação à lei criminal federal, à lei de propriedade intelectual ou à de privacidade das comunicações. 47 U.S.C. § 230(e) (2012). Isso significa que todo provedor de serviços de Internet, mecanismo de busca, plataforma de rede social e site está sujeito a milhares de leis, incluindo leis de pornografia infantil, leis de obscenidade, leis de perseguição e leis de direitos autorais."

[243] FACEBOOK. *Padrões de Comunidade*. Disponível em: <https://transparency.meta.com/pt-br/policies/community-standards/child-sexual-exploitation-abuse-nudity/>. Acesso em 26 dez. 2024: O Facebook define nos "Padrões de Comunidade" a exploração sexual infantil como "[c]onteúdo, atividade ou interações que ameacem, representem, enalteçam, apoiem, forneçam instruções, façam declarações de intenção, admitam a participação ou compartilhem links de exploração sexual de crianças (incluindo menores de idade, crianças pequenas ou bebês reais ou representações não reais com semelhança humana, como em arte, conteúdo gerado por IA, personagens fictícios, bonecos etc.)." Ao estabelecer a proibição de "conteúdo ou atividade que explore crianças sexualmente ou as coloque em perigo", o Facebook também adverte que, embora saibam que "às vezes, as pessoas compartilham imagens dos próprios filhos nus sem más intenções", que essas imagens podem ser removidas "para impedir que outras pessoas cometam abusos, as reutilizem ou se apropriem delas indevidamente".

[244] TOTVS. *Hash: o que é, importância e como funciona*. Publicado em 28 out. 2024. Disponível em: <https://www.totvs.com/blog/gestao-para-assinatura-de-documentos/hash-assinatura-digital/>. Acesso em 26 dez. 2024: "Um hash é uma função criptográfica que transforma uma entrada de dados de qualquer tamanho em uma sequência fixa de caracteres, geralmente composta por números e letras. Essa tecnologia é amplamente utilizada em segurança digital, especialmente na verificação de integridade de dados e em assinaturas digitais. Com ela, é possível garantir que as informações não foram alteradas. No caso da assinatura digital, por exemplo, esse código funciona como uma "impressão digital" que identifica de maneira única um documento. (...) A função hash processa seu conteúdo e cria um código criptografado, único para cada documento."

instantaneamente conteúdos ilegais durante o processo de *upload* do vídeo, para assegurar que imagens nocivas sejam removidas em questão de microssegundos. [245]

O algoritmo *PhotoDNA* se utiliza de uma base de dados que reúne quase 720.000 imagens ilegais disponíveis *online*, organizado pelo *National Center for Missing and Exploited Children* (Centro Nacional para Crianças Desaparecidas e Exploradas) em parceria com o governo estadunidense, para determinar se determinada imagem corresponde ou não um material cadastrado previamente como pornografia infantil.[246]

Possíveis violações de direitos autorais também podem ser moderadas proativamente, antes da publicação, por meio do *software Content ID*. O *software* foi desenvolvido pelo Google, após adquirir em 2006 o *Youtube*, como uma resposta aos anseios da indústria cultural – já que a crescente popularidade do *Youtube* se baseava em grande medida na violação de direitos autorais – e à legislação estadunidense (DMCA), que excepciona a regra de imunidades às plataformas digitais em caso terem conhecimento, por meio do procedimento do *Notice and Takedown*, de violações a direitos autorais.[247]

O *Content ID* permite que detentores de direitos autorais identifiquem seu conteúdo com uma impressão digital (*digital figerprint*), para que possa ser comparado com outros conteúdos postados. O funcionamento do *software* não impede a moderação por meio da sinalização de conteúdos postados (*flagging*) e do regime legal do *Notice and Takedown*. Na realidade, esses sistemas operam conjuntamente. Após o acionamento do sistema de *flagging*, essas informações são adicionadas ao banco de dados do *Content ID* para futura moderação proativa.[248]

Mais do que um mecanismo de filtragem prévia para prevenir violações a direito autorais, o Google transformou o *Content ID* em uma ferramenta de retorno financeiro para os detentores de direitos autorais e, consequentemente, para a própria plataforma. Isso porque, quando identifica um vídeo protegido por direitos autorais, o *Youtube* possibilita imediatamente aos detentores daqueles direitos a opção de *(i)* bloqueá-lo ou *(ii)* incluir propagandas para lhes garantirem um retorno

---

[245] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.637.
[246] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 51.
[247] *Ibidem*, p. 52.
[248] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.637.

financeiro. Esse modelo rendeu bilhões de dólares aos detentores de direitos autorais que escolhem a segunda opção, monetizando o conteúdo de terceiros.[249]

Os incentivos criados pela legislação estadunidense e pelo imperativo de mercado foram fundamentais para criação dessas ferramentas de monitoramento prévio e filtragem algorítmica. A esse respeito, Rodrigo Nitrini discorre da seguinte forma sobre o desenvolvimento do "código":

> Os dois campos mencionados acima – controle prévio e automatizado para combater a pornografia infantil ou a reprodução de não autorizada de conteúdos protegidos por direitos autorais – encontram cada um deles correspondências em incentivos de relevantes leis americanas, que criam nessas searas importantes exceções à regra geral de imunidade jurídica das plataformas em razão [d]e conteúdo produzido ou postado por usuários. Essas leis fixaram obrigações, restrições ou incentivos que terminaram por formatar a criação e a disseminação de tecnologias específicas para a realização dessa filtragem prévia – resultando que, como apontamos, as iniciativas das plataformas foram além daquele que o direito requeria de maneira estrita. Em ambos os casos, imperativos de normas sociais e de mercado tiveram maior peso para o desenvolvimento do 'código'.[250]

A utilização dessas ferramentas foi posteriormente expandida, por iniciativa das próprias plataformas digitais – leia-se: autorregulação –, para outras searas, como o combate à propaganda terrorista. No final do ano de 2016, *Facebook*, *Microsoft*, *X* e *Youtube* anunciaram o compartilhamento de base de dados de impressões digitais (*hashes*) com de imagens e vídeos de terrorismo ou de recrutamento terrorista por elas removidos,[251]-[252] preservando a autonomia das plataformas para realizar a moderação desse conteúdo.[253]

A tendência tecnológica de filtragem algorítmica segue em expansão. O *Facebook* já trabalha para desenvolver a filtragem prévia de vídeos ao vivo, em

---

[249] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 52.
[250] *Ibidem*, p. 53.
[251] *Idem*.
[252] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation*: *upload filters*. Luxembourg: European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, 2020, p. 43. Disponível em: <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf>. Acesso em 26 dez. 2024: No ano de 2020, a base de dados reunidas pelas plataformas no contexto do *Global Internet Forum to Counter Terrorism* (GIFCT) já havia superado mais de 200.000 hashes de vídeos que poderiam ser utilizado para inibir a republicação desses vídeos nas redes sociais.
[253] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 54.

virtude da, cada vez mais frequente, transmissão na plataforma de vídeos de suicídio cometido por jovens. Enquanto a tecnologia ainda não foi implementada, os vídeos ao vivo são moderados primordialmente por meio de denúncias de usuários (*flagging*), sendo possível, por meio da tecnologia existente, que o conteúdo seja marcado, a fim de evitar sua proliferação.[254]

Delineado o funcionamento prático do monitoramento prévio, deve-se tecer breves comentários sobre a sua programação ou, melhor, sobre o seu "código". Os algoritmos não decidem qual tipo de conteúdo deve ser bloqueado antes de ser publicado. O conteúdo filtrado automaticamente geralmente é aquele que pode ser identificado de forma confiável pelo *software* e que é ilegal ou proibido pela plataforma. Esse conjunto de conteúdo que é moderado de forma prévia e automatizada é avaliado e atualizado regularmente por *softwares* iterativos e aprendizado de máquina.[255] Os *softwares* de triagem prévia são, por exemplo, constantemente atualizados pelas redes sociais *Facebook*, *Youtube* e *X* para controlar fontes de *spam* previamente sinalizadas (*flagging*), o que contribui para a melhoria das ferramentas de moderação de conteúdo.[256]

Os sistemas de aprendizagem de máquina (*machine learning*)[257] precisam de uma grande quantidade de dados de treinamento. Os conjuntos de dados utilizados para treinar esses sistemas devem incluir amostras suficientes de cada

---

[254] Foi justamente o que fez o Facebook quando tomou conhecimento da transmissão ao vivo do atentado de Christchurch, ocorrido na Nova Zelândia em março de 2019, e que resultou na morte de 51 pessoas, divulgada em tempo real pelo próprio atirador na plataforma (*Ibidem*, p. 56).

[255] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.637.

[256] *Idem*.

[257] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation…*, p. 37: Existem basicamente três tipos de aprendizagem automática (*machine learning*). O primeiro e mais comum é o aprendizado supervisionado (*supervised learning*), onde a máquina aprende por meio de "supervisão" ou "ensino". Ela recebe um conjunto de treinamento, que é um grande conjunto de respostas corretas para a tarefa do sistema, e aprende a responder a novos casos de forma similar. Por exemplo, um sistema desenvolvido para detectar discurso de ódio ou pornografia pode ser treinado com mensagens que humanos classificaram discurso de ódio ou pornografia. Assim, o sistema aprende a classificar novas mensagens da mesma maneira que no seu treinamento, distinguindo discurso de ofício e pornografia dos demais conteúdos. No aprendizado não supervisionado (*unsupervised learning*), os sistemas aprendem sem receber instruções externas, identificando padrões nos dados. Essas técnicas de aprendizado não supervisionado são particularmente utilizadas para agrupamento de itens que apresentam semelhanças ou conexões relevantes. Por exemplo, documentos que compartilham linguagem ofensiva podem ser agrupados automaticamente. Já no aprendizado por reforço (*reinforcement learning*), a máquina aprende a partir dos resultados de suas próprias ações, observando os resultados e aplicando recompensas ou penalidades (como pontos ganhos ou perdidos) relacionadas a esses resultados. Um exemplo seria um sistema que aprende a priorizar ou não postagens de notícias com base no interesse dos usuários nas postagens apresentadas por ele.

categoria de itens aceitáveis ou inaceitáveis, para que o *software* faça a distinção necessária no momento da moderação. Do contrário, a ferramenta poderá remover conteúdo que não é prejudicial ou manter conteúdo que deveria ser removido.[258]

Para determinados conteúdos, a obtenção de conjuntos de dados adequados não é problemática, especialmente quando as categorias são bem definidas e os critérios de avaliação são claros, como nos casos de violação a direitos autorais ou de pornografia infantil. Essa distinção é mais complexa quando se trata de conteúdos que envolvem questões limítrofes e controvertidas até mesmo para revisores humanos treinados, como, por exemplo, os discursos de ódio. Exemplos dessa complexidade estão fartamente documentos.[259]-[260]

A limitação mais evidente desse modelo é justamente a dependência de base de dados pré-existente que já tenha caracterizado certos conteúdos como ilícitos; os algoritmos têm, por exemplo, dificuldade de conter novos ilícitos. Existem também preocupações quanto à transparência no processo de inclusão de informação nas bases de dados compartilhadas pelas redes sociais.[261]

Enquanto a tecnologia avança sobremaneira, o *design* desses algoritmos permanece em grande parte desconhecido e não regulamentado. Enquanto a literatura das ciências da computação e da comunicação pesquisa sobre a operação e os efeitos desses sistemas complexos, diversos governos estão propondo a sua regulação, com especial ênfase à demanda por maior transparência na utilização dos algoritmos.[262]

2.2.1.2. Controle automatizado de linguagem após a publicação

O controle automatizado é também utilizado para a análise de linguagem posterior à publicação do conteúdo. As plataformas utilizam modelos algorítmicos

---

[258] *Ibidem* p. 45.
[259] *Idem*.
[260] Em março de 2019, as ferramentas do Facebook e do *Youtube* não foram capazes de detectar antes da publicação os vídeos que viralizaram sobre o já mencionado atentado ocorrido em Christchurch, Nova Zelândia. Em contraste, o *YouTube* removeu milhares de vídeos sobre atrocidades na Síria, que poderiam ser utilizados como evidências de crimes de guerra. Mesmo sistemas robustos podem gerar resultados indesejados; o "Content ID" do *YouTube* notificou, em 2018, um músico por violação a direitos autorais após publicar um vídeo de 10 horas com ruído branco (*Ibidem*, p. 46).
[261] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 226.
[262] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*…, p. 34.

que analisam dados em grande escala, buscando padrões linguísticos e contextos que possam indicar a presença de conteúdos ilegais e danosos, como discriminação racial, discurso de ódio, pornografia ou violência.

Um dos desafios dos modelos de linguagem reside na necessidade de operarem em múltiplas línguas, com mesmo grau de eficiência que operariam em língua inglesa. O *Facebook* tem implementado novas abordagens de pré-treinamento *cross-lingual*, desde 2019, para que esses modelos aprendam com a transferência de conhecimento do modelo de linguagem supervisionado, em inglês, e melhorem a sua performance em línguas que têm menos dados disponíveis.[263]

Existem ainda iniciativas para enfrentar conteúdos ilegais contidos em combinações multimídia, com texto e imagem, como, por exemplo, os "memes". Os "memes" são mensagens curtas que combinam texto e imagens ou vídeos, com uma tendência a se reproduzirem rapidamente nas redes sociais. Nas redes sociais, há uma grande variedade de "memes" que veiculam discursos de ódio, baseados em raça, gênero, orientação sexual etc.[264]

O maior desafio para a identificação desses materiais indesejados por mecanismos automatizados é que constituem uma complexa combinação de patrimônio cultural, gírias e expressões idiomáticas que demanda a decodificação de diferentes referências contextuais e socioculturais. Para combater esse problema, o *Facebook* criou um banco de dados com 10.000 "memes" com expressões identificadas como discursos de ódio, disponibilizado para pesquisadores, com o objetivo de permitir o desenvolvimento de métodos para detectar discursos de ódio.[265]

Além disso, o *Facebook* desenvolveu o *Whole Post Integrity Embeddings* ("WPIE"), um *software*, que opera por aprendizado de máquina (*machine learning*), capaz de desempenhar uma análise holística do conteúdo visual e textual de uma postagem e os comentários relacionados, abordando diversas dimensões para verificação de violações às diretrizes da plataforma.[266]

---

[263] META. *Cross-lingual pretraining sets new state of the art for natural language understanding*. Publicado em 4 fev. 2019. Disponível em: <https://ai.meta.com/blog/cross-lingual-pretraining/>. Acesso em 28 dez. 2024.
[264] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation* …, p. 43.
[265] *Idem*.
[266] SCHROEPFER, Mike. *Community standards report*. *Meta*. Publicado em 13 de nov. de 2019. Disponível em: <https://ai.meta.com/blog/community-standards-report/>. Acesso em 28 dez. 2024.

Ao divulgar o Quarto Relatório de Aplicação de Padrões Comunitários (*Fourth Community Standards Enforcement Report*), em novembro de 2019, o *Facebook* explicou o funcionamento do *software* WPIE[267] e apresentou os seus resultados na moderação de conteúdos proibidos pelas suas diretrizes. No terceiro trimestre de 2019, o *Facebook* removeu da plataforma 4,4 milhões de publicações com conteúdos relacionados à venda de drogas, sendo 97,6% desse montante detectados de forma proativa por meio desse mecanismo automatizado.[268]

No Quarto Relatório, o *Facebook* divulgou, ainda, os resultados gerais da utilização de ferramentas automatizadas no primeiro e terceiro trimestres do ano de 2019. No primeiro trimestre de 2019, essas ferramentas foram responsáveis pela remoção, de forma proativa, 4,1 milhões de postagens com discursos de ódio. Já no terceiro trimestre os números são ainda mais expressivos, tendo sido removidos proativamente 7 milhões dessas publicações.[269] De forma semelhante, quase metade de todo o conteúdo removido do *Facebook* ao longo do ano de 2018, com base na

[267] *Idem*: "To get a more holistic understanding of the post, we created Whole Post Integrity Embeddings (WPIE), a pretrained universal representation of content for integrity problems. WPIE works by understanding content across modalities, violation types, and even time. Our latest version is trained on more violations, with greatly increased training data. The system improves performance across modalities by using focal loss, which prevents easy-to-classify examples from overwhelming the detector during training, along with gradient blending, which computes an optimal blend of modalities based on their overfitting behavior.". Tradução: "Para obter uma compreensão mais holística da publicação, criamos o Whole Post Integrity Embeddings (WPIE), uma representação universal pré-treinada de conteúdo para problemas de integridade. O WPIE funciona compreendendo o conteúdo em várias modalidades, tipos de violação e até mesmo no tempo. Nossa versão mais recente é treinada em mais violações, com dados de treinamento muito maiores. O sistema melhora o desempenho em todas as modalidades usando a perda focal, que evita que exemplos fáceis de classificar sobrecarreguem o detector durante o treinamento, juntamente com a combinação de gradientes, que calcula uma combinação ideal de modalidades com base em seu comportamento de superajuste."

[268] *Idem*: "On Facebook, for example, we removed about 4.4 million pieces of drug sale content in Q3 2019, 97.6 percent of which we detected proactively. This is a substantial increase from Q1 2019, when we removed about 841,000 pieces of drug sale content, 84.4 percent of which we detected proactively." Tradução: "No Facebook, por exemplo, removemos cerca de 4,4 milhões de conteúdos de venda de drogas no terceiro trimestre de 2019, 97,6% dos quais detectamos proativamente. Esse é um aumento substancial em relação ao primeiro trimestre de 2019, quando removemos cerca de 841.000 peças de conteúdo de venda de drogas, 84,4% das quais detectamos proativamente".

[269] *Idem*: "This approach is especially helpful for difficult tasks like identifying hate speech because of the nuanced understanding of language that is required. The Community Standards Enforcement Report published today shows how we've improved our proactive detection of hate speech — increasing the amount of content we took action on from 4.1 million pieces in Q1 2019 to 7 million in Q3 2019." Tradução: "Essa abordagem é especialmente útil para tarefas difíceis, como a identificação de discurso de ódio, devido à compreensão diferenciada da linguagem que é necessária. O Relatório de aplicação dos padrões da comunidade publicado hoje mostra como melhoramos nossa detecção proativa de discurso de ódio, aumentando a quantidade de conteúdo sobre o qual tomamos medidas de 4,1 milhões de peças no primeiro trimestre de 2019 para 7 milhões no terceiro trimestre de 2019."

proibição aos discursos de ódio, foi feito proativamente por meio de tecnologias automatizadas (algoritmos), incluindo monitoramento de imagens e textos.[270]

Os dados da moderação de conteúdo revelam que o mecanismo de filtragem algorítmica funciona como uma alternativa viável para o problema do volume massivo das informações compartilhadas nas redes sociais, contrabalanceando a necessidade de atuação permanente de equipes de revisores humanos.[271]

O desenvolvimento e o aperfeiçoamento de controles automatizados foram defendidos por Mark Zuckerberg, em depoimentos prestados ao Congresso estadunidense, em 2018, sobre o escândalo da consultoria *Cambridge Analytica*.[272-273] Nessas audiências, o cofundador do *Facebook* advogou pelo uso de algoritmos como a solução para os desafios da moderação em larga escala: "[a] longo prazo, a criação de ferramentas de inteligência artificial será a maneira escalável de identificar e eliminar a maior parte do conteúdo nocivo [no *Facebook*]".[274]

---

[270] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 57.

[271] *Idem*.

[272] G1. *Cambridge Analytica se declara culpada em caso de uso de dados do Facebook*. Publicado em 9 de jan. de 2019. Disponível: <https://g1.globo.com/economia/tecnologia/noticia/2019/01/09/cambridge-analytica-se-declara-culpada-por-uso-de-dados-do-facebook.ghtml>. Acesso em 27 dez. 2024: "(...) O Facebook já havia admitido que a Cambridge Analytica - uma assessoria política que dirigiu a campanha digital de Trump em 2016 - utilizou um aplicativo para coletar informações privadas de 87 milhões de usuários sem seu conhecimento. A empresa depois utilizou estes dados para mandar aos usuários publicidade política especialmente adaptada e elaborar informes detalhados para ajudar Trump a ganhar a eleição contra a candidata democrata Hillary Clinton".

[273] ARMIJO, Enrique. *Speech Regulation by Algorithm*, 30 Wm. & Mary Bill Rts. J. 245 (2021), pp. 245-263, p. 245. Disponível: <https://scholarship.law.wm.edu/wmborj/vol30/iss2/3>. Acesso em 27 dez. 2024: "Mark Zuckerberg has repeatedly told Congress and other audiences that AI is the key to resolving Facebooks content moderation challenges, envisioning a moderation regime where algorithms detect and take down speech infringing Facebooks Community Standards ex ante, that is, prior to its public posting and before it reaches other users. According to Zuckerberg, this would eventually replace its initial content moderation practices, which relied more on human moderators and user complaints than on automated detection and removal a system that can be slow, inconsistently applied, and often subjects front-line moderators to harrowing emotional harms by exposing them to the worst of the Internet". Tradução: "Mark Zuckerberg afirmou repetidamente ao Congresso e a outros públicos que a inteligência artificial (IA) é a chave para resolver os desafios de moderação de conteúdo do Facebook, imaginando um regime de moderação onde algoritmos detectam e removem conteúdos que infrinjam os Padrões da Comunidade do Facebook ex ante, ou seja, antes de sua publicação e antes de chegarem a outros usuários. Segundo Zuckerberg, isso eventualmente substituiria suas práticas iniciais de moderação de conteúdo, que dependiam mais de moderadores humanos e reclamações dos usuários do que de detecção e remoção automatizadas — um sistema que pode ser lento, aplicado de forma inconsistente e frequentemente expõe os moderadores da linha de frente a danos emocionais profundos ao confrontá-los com o pior da Internet."

[274] C-SAN.GOV. *Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*. Publicado em 10 abr. 2018. Disponível em: <https://www.c-span.org/program/senate-committee/facebook-ceo-mark-zuckerberg-hearing-on-data-privacy-and-protection/500690>. Acesso em 27 dez. 2024.

À medida em que os mecanismos automatizados ganham papel de destaque, preocupações sobre as limitações e os riscos envolvidos no uso dessas tecnologias também crescem. Uma das principais críticas é a de que a inteligência artificial não é capaz de compreender o contexto ou interpretar o verdadeiro significado e a intenção do emissor do discurso.[275]

A dependência dessas tecnologias preocupa ativistas de direitos civis, em razão do baixo desempenho dos algoritmos em decisões que exigem avalições sutis de contexto. É difícil automatizar um algoritmo para distinguir, por exemplo, uma propaganda terrorista de uma matéria jornalística sobre terrorismo. Os filtros algorítmicos também apresentam um desempenho insatisfatório na avaliação de piadas ou sarcasmo ou em idiomas que não são falados por seus desenvolvedores.[276]

Não é possível evitar que erros sejam cometidos pelos mecanismos de filtragem algorítmica baseada em sistemas de aprendizagem de máquina (*machine learning*). O sistema não é infalível porque se baseia em análises probabilísticas. Em um certo nível de desempenho técnico, a redução da taxa de exposição dos usuários a conteúdos que deveriam ser removidos (falsos negativos), geralmente vem acompanhada do indesejado aumento da remoção de conteúdos legais (falsos

---

[275] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 58.

[276] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* …, pp. 33-34: "Reliance on these technologies concerns civil rights activists, since they perform poorly in decisions that require nuanced assessments of context. Distinguishing terrorist propaganda from journalist commentary on terrorism, for instance, or distinguishing content piracy from parody or other fair uses, is difficult to automate. In a 2018 report, the Center for Democracy and Technology reviewed commercially available text-based filters, and found an accuracy rate in the 70-80% range. Filters performed particularly poorly in assessing jokes or sarcasm, or in languages not spoken by their developers. A 2017 report by Princeton Computer Science professor Nick Feamster and Evan Engstrom of the startup-advocacy group Engine provides greater technical detail, analyzing one of the few open-source (and hence publicly reviewable) filtering tools, Echoprint (Engstrom and Feamster 2017). The authors found a 1-2% error rate in simple duplicate matching, including both false positive and false negatives." Tradução: "A dependência dessas tecnologias preocupa ativistas de direitos civis, já que elas têm um desempenho ruim em decisões que exigem avaliações contextuais mais sutis. Distinguir propaganda terrorista de comentários jornalísticos sobre terrorismo, por exemplo, ou diferenciar a pirataria de conteúdo de paródias ou outros usos legítimos, é uma tarefa difícil de automatizar. Em um relatório de 2018, o Center for Democracy and Technology analisou filtros comerciais baseados em texto e identificou uma taxa de precisão na faixa de 70-80%. Os filtros apresentaram desempenho particularmente baixo ao avaliar piadas ou sarcasmo, ou em idiomas que não são falados pelos desenvolvedores. Um relatório de 2017, elaborado pelo professor de Ciência da Computação de Princeton, Nick Feamster, e por Evan Engstrom, da organização de advocacia para startups *Engine*, oferece detalhes técnicos adicionais, analisando uma das poucas ferramentas de filtragem de código aberto (e, portanto, passíveis de revisão pública), o Echoprint (Engstrom and Feamster 2017). Os autores descobriram uma taxa de erro de 1-2% em comparações simples de duplicatas, incluindo falsos positivos e falsos negativos.".

positivos). Em outras palavras, melhorar a sensibilidade pode aumentar a remoção, mas comprometer a precisão.[277]

Além disso, há o risco de que modelos algorítmicos que operam por aprendizagem de máquina (*machine learning*) absorvam vieses discriminatórios já presentes na sociedade ou que estejam incorporados em sua própria criação,[278] causando relevantes alterações no resultado adequado da filtragem ou, melhor, na chamada "verdade fundamental" (*ground truth*).[279]

Na filtragem algorítmica, a *verdade fundamental* não é fornecida por uma realidade física, mas sim por avaliações humanas, de acordo com leis, diretrizes e normas sociais, conforme a interpretação dos avaliadores individuais. Assim, o funcionamento de um sistema de filtragem automatizado pode refletir eventuais preconceitos desses avaliadores humanos cujo comportamento o sistema pretende imitar.[280] A filtragem algorítmica pode, portanto, resultar em discriminação e levar à remoção indevida de conteúdos legais publicados por grupos minorizados.

Nesse sentido, um estudo conduzido pelos pesquisadores do centro de pesquisas "InternetLab" utilizou a ferramenta *Perspective* para analisar o grau de "toxicidade" de mensagens publicadas no X por *drag queens* conhecidas nos Estados Unidos da America em comparação com manifestações de grupos e políticos famosos de extrema direita. A partir das análises feitas pela inteligência artificial, a pesquisa concluiu que "um número significativo de perfis das *drag queens* no Twitter foram considerados como potencialmente mais tóxicos que o perfil de Donald Trump e de supremacistas brancos.".[281]

---

[277] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation*…, p. 45.

[278] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 58.

[279] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation*…, p. 46: "In machine learning domain the term "ground truth" is used to refer to the correct outcome, as identified through standards external to the system, as opposed to the outcome that is proposed by the system. This expression, often used in machine learning apparently derives from cartography, and opposes the representation on a geographical map to the real situation on the ground, which provides the undisputable standard to determine whether the map is correct or wrong." Tradução: "No domínio do aprendizado de máquina, o termo 'verdade fundamental' é usado para se referir ao resultado correto, conforme identificado por meio de padrões externos ao sistema, em oposição ao resultado que é proposto pelo sistema. Essa expressão, usada com frequência no aprendizado de máquina, aparentemente deriva da cartografia e opõe a representação em um mapa geográfico à situação real no local, que fornece o padrão indiscutível para determinar se o mapa está correto ou errado."

[280] *Idem*.

[281] GOMES, Alessandra; ANTONIALLI, Dennys; OLIVA, Thiago. *Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet?*. Disponível em:

O estudo ressaltou que, embora palavras como *bitch*, *fag*, *sissy*, *gay* e *lesbian* possam ser consideradas "tóxicas", conforme concluiu o *software Perspective*, há um número considerável de estudos sobre linguística indicam como o uso de palavras "pseudo-ofensivas" por membros da comunidade LGBTQ desempenham um papel relevante na sua preparação para lidar com a hostilidade externa ao grupo. Ao ignorar o contexto em que estão inseridas essas manifestações, as ferramentas automatizadas podem impactar a capacidade de grupos minorizados de "reclamar esses termos e reforçar vieses danosos".[282]

Outro estudo, conduzido pelo *Center for Democracy and Technology* ("CDT"), detectou vieses discriminatórios na identificação, utilizando-se também da ferramenta *Perspective*, de expressão comuns à população afro-americana.[283] Nas conclusões do CDT, foram listadas ainda limitações na identificação de discursos de ódio: *(i)* melhor funcionamento em determinados contexto; *(ii)* reprodução de vieses discriminatórios contra grupos vulneráveis e marginalizados; *(iii)* necessidade de definição precisa e clara do conteúdo a ser moderado, não podendo ser indicado apenas como "extremismo" ou "radicalismo"; *(iv)* baixo percentual de precisão de acerto, que variam entre 75 e 70%; e *(v)* facilidade de burlar o monitoramento, mediante alteração dos elementos contextuais, podendo ser privilegiados conteúdos sutis, ainda que ilegais.[284]

O controle automatizado, por meio de filtragem algorítmica, é um mecanismo relevante para a curadoria *online*. Esse sistema é capaz de moderar a vasta quantidade de conteúdo compartilhado nas redes sociais, sem intervenção humana imediata, promovendo eficiência e rapidez na moderação de conteúdo. No entanto, também levanta questões sobre a responsabilidade, precisão e neutralidade dessas decisões automatizadas.

<https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/>. Acesso em 28 dez. 2024: "Em média, os níveis de toxicidade das drag queens variam entre 16,68% e 37,81%, enquanto que os níveis dos supremacistas brancos variam entre 21,30% e 28,87%, e o de Trump fica em 21,84%. Também realizamos testes para medir os níveis de toxicidade de palavras usualmente encontradas em tweets de drag queens. A maior parte dessas palavras foram consideradas significativamente tóxicas: BITCH – 98.18%; FAG – 91.94%; SISSY – 83.20%; GAY – 76.10%; LESBIAN – 60.79%; QUEER – 51.03%; TRANSVESTITE – 44.48%. Isso significa que, independentemente do contexto, palavras como "gay", "lesbian" e "queer" já são consideradas como significativamente tóxicas, o que aponta a existência de vieses no Perspective."

[282] *Idem*.

[283] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 59.

[284] *Ibidem*, p. 60.

2.2.1.3. Bloqueio geográfico

O bloqueio geográfico (*geoblocking*) é mais uma forma de controle automatizado e prévio de conteúdo e, portanto, de moderação *ex ante*. Mas ao invés de impedir a publicação de conteúdo ilegal – como é feito, por exemplo, pela *PhotoDNA* (vide subcapítulo 2.2.1.1) –, esse mecanismo envolve tanto o controle prévio à publicação de determinado conteúdo quanto a visualização dele por usuários situados em determinada localização geográfica.[285]

As grandes redes sociais operam globalmente. É natural que as suas políticas de moderação de conteúdo entrem em conflito com normas sociais ou ordenamentos jurídicos de outros países, gerando impasses decorrentes de um "cenário de potencial dinâmica competitiva entre os sistemas normativos das plataformas e das ordens jurídicas tradicionais".[286] É nesse contexto que ganha relevância o mecanismo do bloqueio geográfico, conforme ilustram os exemplos que serão abordados a seguir.

Em novembro de 2006, o governo Tailandês ameaçou bloquear o *Youtube* no país para pressionar o Google a retirar vídeos publicados com conteúdo ofensivo ao Rei Tailandês. Dentre esses vídeos, havia conteúdos que, de fato, violavam os termos de serviço do *Youtube*, mas também imagens satíricas manipuladas, por meio de *Photoshop*, do Rei Tailandês com os pés no lugar da cabeça. Na Tailândia, insultar o Rei é crime com pena de reclusão de até 15 anos.[287]

O *Youtube* enviou, então, à Tailândia a advogada Nicole Wong, especialista na primeira emenda norte-americana – que, dentre outros direitos, assegura a liberdade de expressão –, para resolver a disputa. Após a viagem, Wong defendeu a retirada do conteúdo, após ter tido a oportunidade de observar o amor e a veneração da população tailandesa para com o seu Rei. Assim, o *Youtube* bloqueou o acesso aos vídeos ofensivos ao Rei Tailandês em todo o território tailandês, reconhecendo a violação ao ordenamento jurídico do país.[288]

---

[285] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.637.
[286] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 60.
[287] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.623.
[288] *Idem*.

Poucos meses após esse episódio, uma disputa semelhante emergiu entre o *Youtube* e o governo da Turquia. O desfecho, no entanto, foi ligeiramente diferente. Em março de 2007, a Turquia bloqueou o acesso ao *Youtube* no país, em cumprimento a uma ordem judicial proferida em resposta à publicação de uma paródia de um noticiário que insinuou que o fundador da Turquia moderna, Mustafa Kemal Atatürk, seria *gay*. À semelhança do caso tailandês, também configura crime na Turquia ofensas à Mustafa Kemal Atatürk. Embora o vídeo original tenha sido removido de forma voluntária, a Turquia apresentou ao *Youtube* uma lista de endereços eletrônicos com a republicação do vídeo com a paródia, requerendo a derrubada do conteúdo. O *Youtube* atendeu à demanda da Turquia e promoveu a retirada do conteúdo ofensivo no país. Em junho de 2007, insatisfeita apenas com o bloqueio geográfico, a Turquia exigiu ainda que conteúdo fosse retirado de toda a rede global, o que não foi atendido pela plataforma. Isso levou a Turquia a bloquear o Youtube em seu território.[289]

Outro episódio relevante ocorreu no ano de 2012, desta vez, envolvendo as plataformas do *Facebook*, *Youtube* e *X*, levantando novas questões sobre os riscos à liberdade de expressão por solicitação de diversos governos para retirada do vídeo "Inocência dos Muçulmanos" (*Innocence of Muslims*).[290] O vídeo com o título "A verdadeira vida de Maomé", produzido por uma produtora amadora de filmes situada em Los Angeles, a pretexto de produzir uma biografia visual do profeta – o que por si só já é um tabu religioso para grande dos muçulmanos –[291] representava muçulmanos queimando as casas de cristãos egípcios e a figura do próprio Maomé como um bastardo, homossexual, mulherengo e uma pessoa violenta.[292]

Após a postagem no *Youtube*, o vídeo alcançou milhões de visualizações em menos de um mês. A repercussão do vídeo desencadeou uma onda de protestos violentos que se esparralham pelo mundo – inclusive na Europa, Austrália, Índia e África – e resultaram na morte de 50 pessoas no Paquistão, Afeganistão e Líbia;

---

[289] *Ibidem*, p. 1.624.
[290] *Idem*.
[291] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 32.
[292] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.624.

neste último país, o episódio motivou, ainda, a invasão da embaixada estadunidense no país e a morte de seu embaixador.[293]

O próprio governo americano solicitou publicamente ao *Youtube* a revisão da moderação de conteúdo com base nos termos de uso; o pedido foi negado pela plataforma sob o argumento de que o vídeo estava de acordo com suas políticas de moderação. O *Facebook* também promoveu a revisão do conteúdo e chegou à conclusão semelhante a do *Youtube*, decidindo manter a publicação.[294]

Outros 21 governos nacionais também solicitaram formalmente ao *Youtube* a remoção do vídeo "Inocência dos Muçulmanos". Diante das demandas, o posicionamento inicial da plataforma se alterou, levando a implementação do mecanismo do bloqueio geográfico em países cujo risco de responsabilização parecia mais relevante. Nos países em que possuía uma versão local e representação legal, como Índia, Malásia, Cingapura, Jordânia e Arábia Saudita, o *Youtube* promoveu a remoção do conteúdo. Já no Paquistão, Afeganistão e Bangladesh, onde o *Youtube* não possuía representação local, o pedido foi negado.[295]

Após a escalada de violência na Líbia e no Egito, o *Youtube* anunciou que, embora o conteúdo estivesse conforme as suas regras, restringiria temporariamente acesso ao vídeo nesses países, por meio da ferramenta do bloqueio geográfico, independentemente de solicitação formal por partes de seus governos.[296]

Os casos específicos de ofensas ao Rei tailandês, Mustafa Kemal Atatürk e Profeta Maomé mostram a importância da moderação de conteúdo feita pelas plataformas globais considerar a existência de diferentes contextos sociais e culturais. Afinal, aquilo que é aceitável em um país pode ser inaceitável em outro.[297]

Esses exemplos ilustram também como as plataformas são capazes de conformar suas políticas, modificando-as ou rejeitando eventuais alterações após solicitações formuladas por governos.[298] A possibilidade de as plataformas

---

[293] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 32.
[294] *Idem*.
[295] *Ibidem*, p. 63.
[296] *Idem*.
[297] SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation* …, pp. 46/47.
[298] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.650: "The previous examples of the Thai King, Atatürk, and *Innocence of Muslims* illustrate how platforms have either conformed their policies, modified their policies, or rejected policy changes following government request". Tradução: "Os exemplos anteriores do Rei

moderarem certo conteúdo ou alteração suas políticas internas de moderação por pressão de governos, como já antecipado, gera preocupações sobre o risco de "censura colateral" (*colateral censorship*), cujo conceito pode ser explicado nas palavras de Jack Balkin:

> "A censura colateral ocorre quando o Estado visa responsabilizar 'A' para controlar 'B'. Se 'A' e 'B' forem a mesma empresa ou a mesma publicação, não há problemas significativos para a liberdade de expressão. Por exemplo, consideramos os jornais responsáveis por discursos difamatórios publicações por seus jornalistas, e responsabilizamos as editoras por conteúdo difamatório dos autores que nelas publicam.
> Por outro lado, se 'A' for um provedor de infraestrutura ou canal, como um IPS ou uma rede social, e 'B' for um palestrante independente, então 'A' tenderá a bloquear e a censurar excessivamente para evitar responsabilidade ou sanção governamental. Isso ocorre poque não é o discurso de 'A' que está em jogo, mas o de um estranho, 'B'.[299]

O conceito de *censura colateral* está diretamente relacionado com o regime de responsabilidade das plataformas digitais por conteúdo gerados por terceiros,[300] sendo, portanto, um elemento importante para a avaliação do modelo adequado de regulação das redes sociais.

Assim, o bloqueio geográfico (*geoblocking*) é mais uma ferramenta no arsenal de mecanismo de moderação de conteúdo *ex ante*, mas que gera debates sobre os riscos à liberdade de expressão e de aumento da *censura colateral*, por conta da possibilidade de governos exercerem controle e pressão no contexto da moderação de conteúdo sobre as plataformas digitais.

Tailandês, de Atatürk e da Inocência dos muçulmanos ilustram como as plataformas se conformaram com suas políticas, modificaram suas políticas ou rejeitaram as mudanças de políticas após solicitação do governo".

[299] BALKIN, Jack M., *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation…*, p. 30: "The first key feature of new school speech regulation is collateral censorship. Collateral censorship occurs when the state aims at *A* in order to control *B*'s speech. If *A* and *B* are the same enterprise or the same publication, there is not a significant free speech problem. For example, we hold newspapers liable for defamatory speech published by their reporters, and we hold publishers liable for defamatory content by the authors they publish. On the other hand, if *A* is an infrastructure provider or conduit like an ISP or a social media site, and *B* is an independent speaker, then A will tend to overblock and over-censor to avoid liability or government sanction. That is because it is not *A*'s speech that is at stake, but that of a stranger, *B*."

[300] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 41.

2.2.1.4. Sistema de Flagging

O sistema de *flagging* é um mecanismo *ex post* que permite que os próprios usuários marquem ou sinalizem (*flag*) publicações com a finalidade de denunciar conteúdos potencialmente ofensivos ou ilegais. Após a marcação ou a sinalização, o conteúdo será revisado por moderadores humanos, com base nos termos de uso das plataformas.[301]

O sistema de *flagging* é uma ferramenta que enfrenta diretamente o problema de escala das grandes redes sociais e a impossibilidade de revisão proativa e completa de todo o conteúdo compartilhado nessas plataformas. Por esse motivo, pode-se afirmar que "diante de um desafio (moderação) no qual a escala de publicações se torna um problema em si, o 'flagging', mais do que uma possibilidade operacional, é uma necessidade".[302]

A adoção do *flagging* desempenha duas funções principais: *(i)* oferecer uma forma prática de revisão de volume massivos de conteúdo; e *(ii)* aumentar o grau de legitimidade e confiabilidade da moderação pelos usuários, já que eles próprios são partes fundamentais desse sistema, no qual a revisão está condicionada à prévia sinalização (*flag*) do conteúdo a ser analisado, reduzindo questionamentos sobre a remoção de conteúdo ou censura.[303]

Apesar de volume expressivo de denúncias, que alcançaram mais de um milhão de sinalizações (*flags*) diariamente,[304] a maioria das marcações de conteúdo feitas no *Facebook* não viola os seus Padrões de Comunidade. É muito comum que boa parte dessas denúncias decorreram de conflitos internos de grupo ou divergências de opinião.[305]

Em 2014, um único usuário do *Facebook* marcou centenas de perfis de "drag queens", acusando-as da utilização de nomes falsos, o que levou a plataforma a suspender diversas dessas contas, por violação da política de obrigatoriedade de uso de "nomes reais". Após o assunto vir a público, a partir de diversas reclamações da

---

[301] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.638.
[302] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 64.
[303] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, p. 1.638.
[304] *Idem*: "Facebook users flag over one million pieces of content worldwide every day". Tradução: "Os usuários do Facebook denunciam mais de um milhão de conteúdos em todo o mundo todos os dias".
[305] *Idem*.

comunidade LGBTQIA+, o *Facebook* pediu desculpas e, reinterpretando suas políticas internas, aceitou como "nomes reais" os nomes sociais das "drag queens". Esse episódio evidencia como o sistema de *flagging* pode, por exemplo, ser utilizado por usuários por motivações ideológicas ou políticas.[306]

Para enfrentar essa tendência indesejada, o *Facebook* já implementou um "fluxo de reporte" (*flow*) que leva o usuário a preencher um questionário, para descrever o motivo da denúncia, utilizando termos como "discurso de ódio", "violência ou comportamento prejudicial" ou apenas "não gosto desta publicação". Em caso de assédio ou automutilação, os usuários são direcionados para a opção de "reporte social", que denuncia o conteúdo para a plataforma e para os amigos do usuário. Esse formato de denúncia serve para auxiliar na classificação de conteúdos e na organização da revisão por moderadores humanos, que poderá, assim, priorizar casos mais urgentes, como os de suicídio, violência ou terrorismo.[307]

Após o conteúdo ser marcado, passa-se à fase de revisão por moderadores humanos, que, muitas vezes, operam sob sigilo e com base em diretrizes internas que igualmente não são divulgadas publicamente ou que são constantemente modificadas pelas plataformas digitais.[308]

2.2.1.5. Moderação por revisores humanos

A moderação de conteúdo nas redes sociais não se limita a utilização de ferramentas automatizadas. Após a denúncia realizada pelo sistema de *flagging*, a revisão do conteúdo reportado por moderadores humanos desempenha um papel essencial na moderação (*ex post*) de conteúdo, a partir da interpretação e da aplicação das regras de moderação de conteúdo das redes sociais.

A análise por revisores humanas é fundamental, porque apenas ela permite a real compreensão de elementos contextuais das publicações, podendo, inclusive, ser imprescindível para reavaliação de resultados inadequados decorrentes da utilização de mecanismos automatizados, que operam por filtragem algorítmica e aprendizado de máquina.

---

[306] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 66.
[307] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, p. 1.638.
[308] *Ibidem*, p. 1.639.

O *Facebook* possui cerca de 20 mil moderadores, de diversas nacionalidades e que, muitas vezes, são fluentes em mais de uma língua, possibilitando a moderação de conteúdo em mais de 50 línguas diferentes.[309] A atuação desses moderadores humanos é organizada em formato de pirâmide,[310] que está estruturada e dividida em três níveis de moderadores, conforme ilustração abaixo preparada neste trabalho de pesquisa, com base nas lições de Kate Klonick:[311]

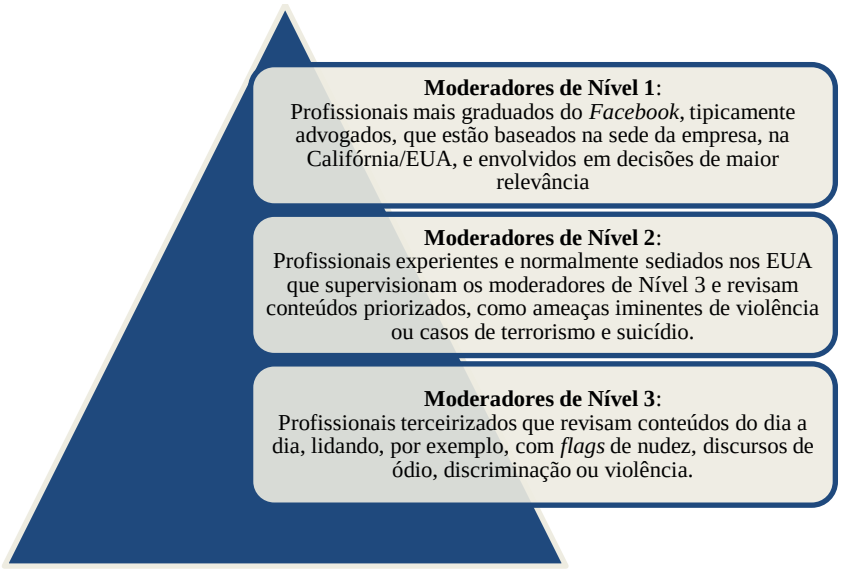

Figura 1 – Pirâmide com os níveis de moderação do Facebook

Nos primeiros anos do *Facebook*, a moderação de nível 3 era desempenhada por jovens recém-formados, em São Franciso. Atualmente, boa parte desse nível de moderação é feita por profissionais terceirizados, que trabalham em *call centers* espalhados por todo o mundo – Filipinas, Irlanda, México, Turquia, Índia, Leste Europeu e nos próprios Estados Unidos da América.[312]

Após a classificação do conteúdo reportado ("flag"), com a definição da prioridade de moderação, os moderadores do nível 3 iniciam a revisão dos

---

[309] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 67.
[310] *Ibidem*, p. 68.
[311] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, p. 1.639-1.640: "At Facebook, there are three basic tiers of content moderators: "Tier 3" moderators, who do the majority of the day-to-day reviewing of content; "Tier 2" moderators, who supervise Tier 3 moderators and review prioritized or escalated content; and "Tier 1" moderators, who are typically lawyers or policymakers based at company headquarters". Tradução: "No Facebook, há três níveis básicos de moderadores de conteúdo: Moderadores de 'Nível 3', que fazem a maior parte da revisão diária do conteúdo; moderadores de 'Nível 2', que supervisionam os moderadores de Nível 3 e revisam o conteúdo priorizado ou escalado; e moderadores de 'Nível 1', que normalmente são advogados ou formuladores de políticas baseados na sede da empresa".
[312] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.640.

conteúdos do dia a dia. O mesmo conteúdo é submetido a mais de um moderador do nível 3 para verificação de eventuais divergências nas decisões de moderação. Caso confirmada a divergência, o conteúdo é escalonado para um moderador do nível 2 para uma nova avaliação.[313]

Os moderadores de nível 2 revisam os conteúdos que foram priorizados de acordo com as indicações da denúncia pelo sistema de *flagging* ou, como visto acima, os conteúdos revisados por mais de um moderador do nível 3 que resultaram em decisões de moderação divergentes. Os moderadores de nível 1, por sua vez, são responsáveis pelas decisões mais relevantes, que são tomadas na sede da empresa.[314]

A tarefa dos moderadores humanos envolve duas importantes análises que devem ser desempenhadas de forma sucessiva. Primeiro, o moderador deve identificar se há violação às regras da plataforma. Em seguida, se houver violação, o revisor precisa especificar qual a regra foi violada. Se houver qualquer equívoco na indicação da regra violada, o resultado não será contabilizado como acerto, ainda que a decisão de moderação tenha sido correta.[315]

A respeito da dinâmica de moderação por revisores humanos, Kate Klonick defende que "moderadores são treinados para aplicar as normas internas ou 'Políticas de Abuso' ['*Abuse Standards*'], com um enfoque que imita a jurisprudência moderna", comparando-os com "juízes", que utilizam "conceitos legais" e realizam "análises que envolvem analogias e testes multifatoriais para avaliar se o conteúdo sinalizado viola as normas da plataforma".[316-317]

A aplicação das normas de moderação também inclui a possibilidade do usuário que tiver seu conteúdo moderado apresentar recursos contra as decisões dos

---

[313] *Ibidem*, p. 1.641.
[314] *Idem*.
[315] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas...*, p. 69.
[316] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, pp. 1.641-1.642.
[317] Rodrigo Nitrini critica a abordagem de Kate Klonick ao comparar os moderadores com juízes. Para o autor, a comparação seria falha, porque os moderadores humanos são empregados das empresas de tecnologia e devem decidir necessariamente de acordo com regras pré-estabelecidas, com pouquíssimo ou nenhum poder para superar qualquer inconsistência, por mais que ela pareça necessária de acordo com o contexto fático. Esse papel quase automatizado no processo de aplicação das regras de moderação se distinguiria daquele esperado dos juízos, que precisam "articular as razões de suas decisões com razoável independência". (NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., p. 71).

moderadores humanos. No *Facebook*,[318] os recursos podem ser apresentados nos casos de remoção de conteúdo, suspensão ou banimento de perfis ou páginas, embora não esteja disponível para qualquer decisão baseada em violação às suas "Padrões de Comunidade"; a plataforma pune usuários que têm conteúdo removido repetidamente com suspensões que podem começar em 24 horas e aumentar gradualmente até o banimento.[319] No *Youtube*,[320] os recursos podem ser apresentados contra remoção de conteúdo e em caso de aplicação de outras penalidades. Já no *X*, qualquer decisão tomada pela plataforma com base nas suas regras e políticas internas pode ser contestada por usuários.[321_322]

A revisão humana possibilita a revisão de decisões equivocadas de moderadores individuais, por meio de mecanismos de redundância, assim como a correção de vieses discriminatórios expressos em resultados indesejados do controle automatizado feito por modelos algorítmicos. Dessa forma, a moderação por revisores humanos, após o acionamento do mecanismo do *flagging*, é um componente crucial na moderação de conteúdo das redes sociais, em complemento à utilização de ferramentas automatizadas, contribuindo para o aperfeiçoamento do sistema de aplicação das regras das plataformas.

2.3. Limites e desafios da autorregulação na aplicação dos termos de uso das redes sociais

As redes sociais exercem o controle do exercício da liberdade de expressão, quer quando o restringem, limitando o acesso ou removendo conteúdo, bem como

---

[318] Ao longo do ano de 2019, o Facebook analisou cerca de 15 milhões de recursos contra decisões tomadas com base nas suas regras. Dentre esses recursos, pouco mais de 3 milhões recursos foram providos, o que representa um percentual de êxito recursal de 20%, conforme informações extraídas de seu relatório de transparência (SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation* …, p 50).

[319] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*…, p. 1.647

[320] Em seu relatório de transparência, o *YouTube* informou que, no primeiro trimestre de 2020, de um total de 6.111.008 vídeos que foram removidos por violarem suas regras, cerca de 2% dessas decisões de moderação foram objeto de recurso (165.941 recursos). Esses recursos foram analisados por moderadores experientes, que reverteram quase 25% das decisões recorridas, restabelecendo um total de 41.059 vídeos na plataforma. (SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation*…, p 50).

[321] *Idem*: No *X*, o procedimento de contestação de eventual suspensão ou bloqueio de conta é normalmente finalizado em até 7 dias. Esse procedimento, no entanto, carece de transparência. O *X* não divulgar em seus relatórios de transparência o número de recursos apresentados com as decisões de moderação tampouco o número de contas e mensagens reintegrados após a análise dos recursos.

[322] KLONICK, Kate. *The New Governors: The People, Rules, and Processes Governing Online Speech*, … p. 1.648.

suspendendo ou excluindo contas e perfis de usuários, quer quando nada fazem, permitindo que certo conteúdo danoso ou ilegal seja disseminado, em detrimento dos direitos de seus usuários.[323]

Esse poder das plataformas digitais está justificado, conforme desenvolvido acima, em virtude do exercício da autonomia privada e da liberdade editorial – que irradia da liberdade de expressão – das plataformas. Esses direitos garantem às plataformas a livre estipulação e aplicação de seus termos de uso ou padrões/diretrizes de comunidade,[324] para definição de que tipo de ambiente pretendem proporcionar aos seus usuários.

Luna Barroso ressalta que "a liberdade para definir o tipo de comunidade que pretendem criar na internet é um componente essencial da livre iniciativa e da liberdade de expressão dessas empresas". Essa liberdade, segundo a autora, está condicionada, contudo, ao fornecimento de "informações necessárias e suficientes para que os usuários entendam, a partir de termos de uso claros e específicos, o que é proibido",[325] que visam a corrigir a assimetria informacional e o déficit de conhecimento dos usuários.

Nesse contexto, vislumbra-se a existência de limites negativos à autorregulação, que decorrem dos direitos fundamentais dos usuários, notadamente, da liberdade de expressão, do direito à privacidade[326], do direito à honra[327] e do direito à proteção de seus dados pessoais. Afinal, são esses os principais direitos que podem ser postos em causa quando os usuários não conhecem de forma clara as regras de autorregulação. Essa conclusão adianta uma relevante discussão sobre a eficácia horizontal dos direitos fundamentais nas relações privadas entre plataformas digitais e usuários, que será abordada adiante.

As ponderações feitas pelas redes sociais na moderação de conteúdo ao impor limitações às liberdades fundamentais de seus usuários revelam que é necessário que as plataformas digitais observem mandamentos procedimentais de *transparência* e *accountability* (prestação de contas). O estado da arte aponta que

---

[323] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., p. 43-104, pp. 74-75.
[324] *Ibidem*, p. 73.
[325] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 222.
[326] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., pp. 43-104, p. 73.
[327] *Ibidem*, p. 81.

esses são desafios da autorregulação, cujo atingimento é fundamental para garantir a legitimidade de sua atuação na curadoria do conteúdo *online*.

Quando se fala em limites ou desafios da autorregulação na aplicação dos termos de uso das redes sociais, pode-se também pensar em limitações de ordem prática da moderação de conteúdo. Conforme já explicado, a moderação por revisores humanos é colocada à prova pela impossibilidade técnica de revisão da massiva escala de conteúdos compartilhados. Já a moderação por modelos algorítmicos, embora essencial para a revisão do volume expressivo de publicações nas redes sociais, produz resultados imprecisos, pela dificuldade dos algoritmos em identificar elementos contextuais ou pela sua propensão à absorção de vieses discriminatórios. Essas questões já foram abordadas no subcapítulo 2.2 *supra* e não será preciso revisitá-las.

Existem ainda outros limites e desafios inerentes à autorregulação que estão relacionados aos sistemas de recomendação algorítmica para modulação da experiência de navegação dos usuários e ao modelo de publicidade *online* adotados pelas plataformas digitais. Esses limites e desafios também não serão aqui analisados, por não se referirem especificamente à moderação de conteúdo.

Feito esses esclarecimentos, passa-se a demonstrar os limites negativos dos direitos fundamentais dos usuários à autorregulação na aplicação dos termos de uso, levando-se em consideração a eficácia horizontal desses direitos sobre as relações privadas no ambiente virtual. Em seguida, serão desenvolvidos os desafios de *transparência* e *accountability* das redes sociais na moderação de conteúdo, como meio para assegurar a legitimidade decisória das plataformas.

2.3.1. Limites negativos: eficácia horizontal dos direitos fundamentais dos usuários

Os direitos à autonomia e à liberdade das plataformas digitais para formular e aplicar as regras de moderação de conteúdo, que são previamente aceitas pelos usuários, encontra limites negativos nos direitos à liberdade de expressão, privacidade e segurança de seus usuários. Esses limites remetem à discussão sobre a eficácia horizontal dos direitos fundamentais nas relações privadas como questão central do espectro regulatório das redes sociais.

Os direitos fundamentais foram concebidos historicamente como proteções aos indivíduos contra o abuso de poder estatal, aplicando-se, exclusivamente, na relação (vertical) entre indivíduo e Estado. A percepção de que abusos não estão inseridas apenas nas relações com o Estado, mas também na relação (horizontal) entre particulares, ensejou discussões sobre *se* e *como* os direitos fundamentais se aplicariam nas relações privadas.[328]

O precedente inaugural da teoria da eficácia horizontal dos direitos fundamentais remonta aos idos de 1958 no julgamento do caso Lüth.[329] O precedente do Tribunal Constitucional Federal alemão foi o primeiro a superar, no dizer de Luis Roberto Barroso, "a rigidez da dualidade público-privado ao admitir a aplicação da Constituição às relações particulares, inicialmente regidas pelo Código Civil".[330]

A atual dogmática dos direitos fundamentais aponta para três correntes doutrinárias sobre a eficácia horizontal. A primeira corrente defende que os direitos fundamentais somente se aplicariam nas relações entre indivíduos e Estado.[331] A teoria que nega a eficácia horizontal dos direitos fundamentais nas relações privadas, admitindo apenas a eficácia vertical (indivíduo-Estado), não conhece atualmente seguidores significativos.[332]

A maioria da doutrina admite a produção de efeitos dos direitos fundamentais também nas relações horizontais entre particulares, nas quais o Estado não participa. O cerne da questão não é, portanto, *se* esses direitos produzem

---

[328] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 239.

[329] Erich Lüth, presidente do Clube de Imprensa de Hamburgo, incentivou o boicote a um filme dirigido por Veit Harlan, um cineasta que fora associado no passado ao regime nazista. A produtora e a distribuidora do filme obtiveram, na esfera da justiça ordinária, uma decisão judicial que determinou a cessação do boicote, com base em norma do Código Civil alemão (BGB), que prevê que quem causar danos a outrem de forma contrária aos bons costumes é obrigado a reparar esses danos. No entanto, o Tribunal Constitucional Federal alemão reformou a decisão, respaldando o direito fundamental à liberdade de expressão, que deveria guiar a interpretação do Código Civil alemão. (BARROSO, Luís Roberto. Neoconstitucionalismo e Constitucionalização do Direito: o triunfo tardio do Direito Constitucional no Brasil. *Revista da Escola de Magistratura do Estado do Rio de Janeiro - EMERJ*, v. 9, nº 33, 2006, p. 43-92, p. 61. Disponível em: <https://www.emerj.tjrj.jus.br/revistaemerj_online/edicoes/revista33/Revista33_43.pdf>. Acesso em 5 jan. 2025.

[330] *Ibidem*, p. 75.

[331] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 239.

[332] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., pp. 43-104, p. 61.

efeitos nas relações horizontais, mas *como* esses efeitos são produzidos.[333] Assim, emergem outras duas teorias que partem de uma base comum, mas divergem quanto a forma de aplicação desses direitos nas relações privadas.

A segunda corrente doutrinária entende que os direitos fundamentais devem ser aplicados de forma indireta às relações entre particulares, por meio da reinterpretação de normas de direito infraconstitucional. Finalmente, a terceira corrente entende que os fundamentais devem ser aplicados de forma direta nas relações privadas, da mesma forma como se aplicam às relações entre indivíduos e Estado (relação vertical).[334]

Apesar da controvérsia doutrinária, o Supremo Tribunal Federal já se debruçou sobre a questão em mais de uma ocasião, para concluir que os direitos e garantias fundamentais produzem efeitos nos planos vertical (indivíduo-Estado) e horizontal (indivíduo-indivíduo). A Corte Suprema decidiu, por exemplo, pela invalidade da exclusão de sócio de uma associação, tendo em vista que "os direitos fundamentais assegurados pela Constituição vinculam diretamente não apenas os poderes públicos, estando direcionados também à proteção dos particulares em face dos poderes privados".[335] Mais especificamente sobre a possibilidade de limitação da autonomia privada, o Plenário do Supremo Tribunal Federal concluiu que:

> A autonomia privada, que encontra claras limitações de ordem jurídica, não pode ser exercida em detrimento ou com desrespeito aos direitos e garantias de terceiros, especialmente aqueles positivados em sede constitucional, pois a autonomia da vontade não confere aos particulares, no domínio de sua incidência e atuação, o poder de transgredir ou de ignorar as restrições postas e definidas pela própria Constituição, cuja eficácia e força normativa também se impõem, aos particulares, no âmbito de suas relações privadas, em tema de liberdades fundamentais.[336]

A aplicação imediata e a eficácia das normas constitucionais que estipulam direitos e garantias individuais, prevista no §1° do artigo 5° da Constituição Federal

---

[333] SILVA, Virgílio Afonso da. Direitos fundamentais e relações entre particulares. *Revista Direito GV*, v.1, n. 1, p. 173-180, maio, 2005, p. 174. Disponível em: <https://periodicos.fgv.br/revdireitogv/article/view/35274/34067>. Acesso em 5 jan. 2025.
[334] *Ibidem*, pp. 174-175.
[335] BRASIL. Supremo Tribunal Federal. RE n. 201819/RJ. Minª. Relª. Ellen Gracie, Rel. p/ Acórdão Min. Gilmar Mendes, 2ª Turma, j. 11 out. 2005.
[336] BRASIL. Supremo Tribunal Federal. RE n. 639138/RS. Min. Gilmar Mendes, Rel. p/ Acórdão Min. Edson Fachin, Plenário, j. 18 ago. 2020 (Tema 452/STF).

de 1988,[337] atuam, conforme explica Alexandre de Moraes, "em dois planos distintos e complementares – eficácia vertical (Estado-indivíduo) e horizontal (indivíduo-indivíduo), de maneira a evitar abusos e excessos inconstitucionais tanto na atuação estatal quanto nas relações privadas e sociais".[338]

O posicionamento da jurisprudência do Supremo Tribunal Federal sobre a eficácia horizontal dos direitos fundamentais fortalece o crescente fenômeno do chamado "Constitucionalismo Digital". A teoria do Constitucionalismo Digital surgiu incialmente como um movimento constitucional de defesa da limitação do poder de atores privados da internet, como as plataformas digitais. Mais recentemente, o fenômeno tem abrangido iniciativas jurídicas e políticas, estatais e não estatais, relacionadas à afirmação de direitos fundamentais na internet (*Internet Bill of Rights*).[339]

A doutrina diverge quanto ao conceito e ao alcance desse fenômeno. O Constitucionalismo Digital pode ser compreendido ora como uma corrente teórica do Direito Constitucional contemporâneo que se organiza a partir do estabelecimento de normas que reconhecem, afirmam e garantem direitos fundamentais na internet,[340] ora como uma ideologia constitucional que visa a estabelecer e garantir uma estrutura normativa para a proteção dos direitos fundamentais e o equilíbrio de poderes no ambiente virtual.[341]

---

[337] Constituição de 1988: "Artigo 5°. (...) § 1º As normas definidoras dos direitos e garantias fundamentais têm aplicação imediata."

[338] MORAES, Alexandre de. *Direito Constitucional*. 40ª Edição, 2024. Rio de Janeiro: Atlas, 2024, *ebook*. p. 36.

[339] MENDES, Gilmar Ferreira; FERNANDES, Victor Oliveira. Constitucionalismo digital e jurisdição constitucional: uma agenda de pesquisa para o caso brasileiro. *Revista Brasileira de Direito*, Passo Fundo, vol. 16, n. 1, pp. 1-33, jan.-abr., 2020, pp. 4-5.

[340] *Ibidem*, p. 5: "Para os fins do presente estudo, entende-se que o Constitucionalismo Digital corresponde, de forma ainda mais abstrata, a uma corrente teórica do Direito Constitucional contemporâneo que se organiza a partir de prescrições normativas comuns de reconhecimento, afirmação e proteção de direitos fundamentais no ciberespaço".

[341] CELESTE, Edoardo. Digital Constitutionalism: Mapping the Constitutional Response to Digital Technology's Challenges. *HIIG Discussion Paper Series*, n. 02, 2018, p. 15. Disponível em: <https://ssrn.com/abstract=3219905>. Acesso em 5 jan. 2025: "I consider digital constitutionalism as a declination of modern constitutionalism. The former shares the foundational values, the overall aims of the latter, but it focuses on the specific context affected by the advent of digital technology. Being digital constitutionalism an ism, one could define it as the ideology which aims to establish and to ensure the existence of a normative framework for the protection of fundamental rights and the balancing of powers in the digital environment.". Tradução: "Considero o constitucionalismo digital como um desdobramento do constitucionalismo moderno. O primeiro compartilha os valores fundamentais e os objetivos gerais do segundo, mas se concentra no contexto específico afetado pelo advento da tecnologia digital. Sendo o constitucionalismo digital um "ismo", pode-se defini-lo como a ideologia que visa estabelecer e garantir a existência de uma estrutura normativa para a proteção dos direitos fundamentais e o equilíbrio de poderes no ambiente digital."

O tema do Constitucionalismo Digital envolve complexidades próprias e não será aprofundado aqui. Para os propósitos deste estudo, pode-se concluir, com base nos ensinamentos doutrinários acima mencionados, que prevalece, no Constitucionalismo Digital, a ideia de que os direitos e garantias fundamentais previstos nas normas constitucionais produzem efeitos nas relações entre os usuários e as plataformas privadas.

O necessário reconhecimento da eficácia horizontal dos direitos fundamentais nas relações privadas no ambiente virtual consiste em uma limitação negativa da autorregulação na estipulação e aplicação dos termos de uso das plataformas. Isso não significa, contudo, que exista uma vedação para as redes sociais realizarem a moderação de conteúdo com base em regras privadas.

A moderação de conteúdo nas redes sociais constitui uma forma de sanção contratual previamente prevista nos termos de uso e cuja aplicação forçada é feita pelas próprias plataformas digitais, o que poderia ser interpretado como uma espécie de autotutela admitida pelo ordenamento jurídico, conforme explicado no subcapítulo 2.1.1 *supra*. Além disso, é fundamental que seja preservada a autonomia das plataformas para determinar o tipo de comunidade que desejam oferecer aos seus usuários, garantindo o exercício dos direitos à liberdade de iniciativa e à liberdade de expressão dessas empresas.

Os limites negativos que decorrem da eficácia dos direitos fundamentais sobre as relações no ambiente virtual, ao revés, impõem a observância de deveres procedimentais específicos que, muitas vezes, não são adequadamente atendidos pelas redes sociais, como, por exemplo, de transparência e *accountability*. Esses efeitos poderão, por outro lado, criar também limites positivos para uma intervenção regulatória pública que vise a atender a interesses públicos relevantes, conforme será abordado no subcapítulo 3.3 *infra*.

2.3.2. Desafios de transparência e *accountability*

Alguns dos aspectos mais criticados das redes sociais dizem respeito à aplicação dos termos de uso das plataformas, em especial pela opacidade dos algoritmos utilizados na moderação de conteúdo e pela falta de clareza das regras que embasam o controle exercido sobre a liberdade de expressão dos usuários. Por esse motivo, há certo consenso entre os estudiosos de que as plataformas devem

conferir maior transparência às decisões tomadas na moderação de conteúdo. Afinal, transparência é um pressuposto essencial ao exercício de qualquer outro direito procedimental, como o devido processo e a isonomia.[342]

A garantia de maior transparência contribui para a melhoria do debate público sobre a aplicação das políticas internas e as decisões adotadas pelas plataformas digitais na moderação de conteúdo, permitindo maior controle da sociedade sobre a sua atuação.[343] A compreensão dos critérios que dão suporte às decisões das plataformas digitais na identificação, análise e retirada de conteúdos danosos ou ilegais e dos parâmetros utilizados pelos modelos algorítmicos é também crucial para que essas decisões sejam incrementadas do ponto de vista de legitimidade democrática.[344]

A facilitação do acesso a dados a pesquisadores e estudiosos, por sua vez, é uma importante medida para a compreensão de questões caras à democracia, como, por exemplo, o papel de agentes estatais e não estatais na manipulação de processos eleitorais, bem como de modelos algorítmicos no impulsionamento e na ampliação de informações falsas e ataques democráticos.[345] Medidas de transparência são comumente associadas, portanto, a diversos objetivos específicos e complementares, trazendo benefícios na identificação de problemas complexos e na busca por soluções adequadas relacionadas à moderação de conteúdo realizada pelas plataformas, sem suscitar, por outro lado, maiores riscos à liberdade de expressão dos usuários.

2.3.1.1. Objetivos

O cumprimento de regras de transparência pelas redes sociais pode servir, como mencionado, a diversos objetivos específicos e complementares. O primeiro deles é o de garantir que usuários, na qualidade de consumidores, tenham capacidade de tomar decisões informadas. O fornecimento de informações claras e precisas cumpre a finalidade de empoderar os usuários a tomar decisões conscientes

---

[342] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 242.
[343] *Ibidem*, pp. 242-243.
[344] KELLER, Clara Iglesias; MENDES, Laura Schertel; FERNANDES, Victor. *Moderação de conteúdo em plataformas digitais: caminhos para a regulação no Brasil*..., pp. 79-80.
[345] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 243.

sobre quais bens e serviços consideram atraentes ao seu consumo. A transparência está intrinsicamente vinculada à proteção do consumidor, cumprindo, assim, o papel de coibir a adoção de práticas desleais e enganosas.[346]

No contexto da moderação de conteúdo, essas informações dizem respeito ao detalhamento de regras e procedimentos adotados pelas plataformas. Incluem a regular publicação de relatórios sobre o funcionamento de seus sistemas de moderação de conteúdo, bem como a permissão de acesso externo aos dados da plataforma por parte de pesquisadores e legisladores para a condução de auditorias. Assim, o público consumidor, uma vez informado sobre o funcionamento desses sistemas, poderá avaliar se essas operações atendem, por exemplo, ao interesse público.[347]

As regras de transparência cumprem uma função relacionada à formação da opinião pública sobre as operações e os sistemas de moderação das plataformas. Essas empresas podem ter interesse em adotar uma postura de negação quanto aos efeitos adversos da operação de seus sistemas de moderação de conteúdo. No entanto, o fornecimento de informações ao público, estudiosos e especialistas possibilita que estudos, auditorias e avaliações externas sejam conduzidas de forma independente, para embasar a opinião pública.[348]

Assim, mesmo que as plataformas optem por não identificar e endereçar esses problemas, pesquisadores independentes poderão preencher essa lacuna divulgando seus estudos ao público, que, por sua vez, poderá pressionar por mudanças em seus sistemas de moderação de conteúdo. Essa dinâmica cria uma interativa política pública de diálogo, com capacidade de escalar para soluções de melhorias estruturais baseadas em *feedback*. Essas melhorias podem ser aplicadas tanto para as próprias medidas de transparência, quanto para mandamentos mais amplos, como os chamados deveres de cuidados (*duty of care*).[349]

A transparência também é um elemento chave para possibilitar o cumprimento de medidas de *accountability* pelas plataformas digitais. A ampla divulgação de informações contribuem para a adoção de medidas de melhoria dos

---

[346] MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry* (February 12, 2020). Transatlantic Working Group, 2020, p. 7. Disponível em: <https://ssrn.com/abstract=3615726>. Acesso em 11 jan. 2025.
[347] *Idem*.
[348] *Ibidem*, p. 8.
[349] *Idem.*

sistemas de moderação de conteúdo recomendadas por estudiosos do tema. Alguns exemplos envolvem a implementação de *(i)* uma corte especial da *internet* que utilize normas locais (ao invés de regras de termos de uso ou de diretrizes de comunidade) para agilizar o julgamento de remoção de determinados conteúdos; *(ii)* conselhos independentes que poderiam endereçar e supervisionar as práticas de moderação de conteúdo das grandes plataformas digitais; e *(iii)* um procedimento de revisão de solicitação de remoção de conteúdo não atendidas, nos moldes daquelas concedidas àqueles que tiveram conteúdos moderados.[350]

A recomendação de estabelecimento de um conselho de supervisão independente para a análise de decisões mais críticas de moderação de conteúdo, foi, inclusive, implementada pelo *Facebook*. Em 2008, o *Facebook* criou um órgão independente, formado por pessoas não vinculadas à empresa, para tomar decisões finais, transparentes e vinculantes no campo da moderação de conteúdo, chamado de *Oversight Board* ("Comitê Supervisor"). O Comitê Supervisor se dedica à supervisão da moderação de casos relevantes e paradigmáticos, tendo autonomia e liberdade para a escolha de quantos e quais casos serão julgados, proferindo decisões com caráter vinculativo para o caso analisado. O Comitê possui atribuição para sugerir mudanças das regras da política de moderação de conteúdo, sobre as quais o *Facebook* deve se manifestar, aceitando a sugestão e promovendo a mudança recomendada ou recusando-se e, nesse caso, fornecendo as razões para a rejeição.[351-352]

---

[350] *Ibidem*, p. 7.

[351] NITRINI, Rodrigo Vidal. *Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas*..., pp. 143-144.

[352] Em caso interessante, o Comitê Supervisor avaliou alguns vídeos sobre os atentados do 8 de janeiro de 2024 após um usuário apelar contra a decisão da Meta de entender que os vídeos não violariam seus termos de uso. No dia 3 de janeiro de 2023, um usuário postou no *Facebook* um vídeo sobre as eleições brasileiras, defendendo o cerco ao Congresso Nacional como "última alternativa". O vídeo exibia um general fardado, apoiador da reeleição de Jair Bolsonaro, incentivando as pessoas a "tomar as ruas" e a "ir ao Congresso Nacional e ao Supremo Tribunal Federal". Em seguida, apareciam imagens de um incêndio na Praça dos Três Poderes, com textos sobrepostos: "Venha para Brasília! Vamos invadir! Vamos cercar os três poderes." Outra imagem trazia a frase "exigimos o código-fonte", usada por manifestantes para contestar a confiabilidade das urnas eletrônicas. No mesmo dia, um usuário denunciou o conteúdo por violar as Diretrizes da Comunidade da Meta sobre Violência e Incitação, que proíbem chamados para invasão forçada de locais de alto risco. Entre os dias 3 e 4 de janeiro de 2024, quatro usuários denunciaram o conteúdo sete vezes. A Meta analisou a primeira denúncia, mas concluiu que o vídeo não violava suas políticas. O usuário recorreu, mas a decisão foi mantida. Nos dias seguintes, outras seis denúncias foram analisadas por cinco moderadores diferentes, que também não consideraram o conteúdo violador. Em 8 de janeiro de 2024, apoiadores do ex-presidente Bolsonaro invadiram o Congresso Nacional, o STF e o Palácio do Planalto, destruindo propriedades e confrontando a polícia. No dia seguinte, a Meta classificou

Luna Barroso afirma que a transparência serve 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 um debate qualificado que busque aprimorar as práticas da indústria como um todo". A autora elenca outros dois objetivos complementares da garantia de transparência. Primeiro, fornecer aos usuários maior compreensão e conhecimento sobre a atuação das plataformas na regulação do discurso público, permitindo que se mantenham seguros e prevenindo danos no ambiente digital. E, ainda, garantir ao judiciário ou ao órgão regulador designado, e a pesquisadores, maiores informações para compreensão das ameaças dos serviços digitais, o papel das plataformas na minimização ou amplificação desses riscos, e eventuais ações de mitigação de danos adotadas.[353]

Reconhecendo a proteção do consumidor, por meio do fornecimento de informações necessárias para que possam tomar decisões informadas na escolha de bens e serviços, e um objetivo regulatório, através de exigências para que as plataformas divulguem informações essenciais à aplicação de outras leis, Daphne Keller defende que o principal objetivo das regras de transparência é o "democrático", relacionado ao compartilhamento de informações que legisladores

---

os ataques como um "evento violador" sob sua política de Pessoas e Organizações Perigosas, anunciando a remoção de conteúdos que apoiassem ou elogiassem as ações. A empresa também declarou o Brasil uma "localização de alto risco temporária" e começou a remover postagens que incentivassem a invasão de prédios federais. Após a seleção do caso pelo Comitê Supervisor, a Meta reconheceu seu erro ao manter o vídeo no ar e o removeu em 20 de janeiro. O Comitê avaliou que os esforços da Meta para garantir a integridade eleitoral no Brasil e em outros países são preocupantes. Embora questionar eleições seja, em geral, um direito protegido, alegações amplamente disseminadas podem levar à violência em certos contextos. No caso, a intenção do autor, o conteúdo do discurso, seu alcance e o risco iminente de danos justificavam a remoção da postagem. Como o post chamava explicitamente para a invasão de prédios governamentais na Praça dos Três Poderes — que se encaixam nessas categorias —, a decisão da Meta de manter o conteúdo no ar, em meio a uma crise política, contrariou suas próprias regras. O Comitê demonstrou preocupação com o fato de que, apesar da instabilidade no Brasil e da disseminação de conteúdos semelhantes antes do ataque de 8 de janeiro, os moderadores da Meta não identificaram a publicação como violadora nem escalaram o caso para revisão mais detalhada. O conteúdo só foi removido duas semanas depois, quando o evento ao qual se referia já havia ocorrido, e apenas após o Comitê intervir. Diante disso, o Comitê reverteu a decisão da Meta de manter a publicação no ar e recomendou que a empresa: (i) desenvolva um modelo para avaliar seus esforços de integridade eleitoral, incluindo métricas sobre a aplicação de suas políticas de conteúdo e a moderação de anúncios; e (ii) esclareça em seu Centro de Transparência que, além do Protocolo de Política de Crise, adota outros protocolos para prevenir e lidar com riscos em contextos eleitorais e eventos de alto risco. (OVERSIGHT BOARD. *Discurso do General Brasileiro*. Publicado em 22 jun. 2023. Disponível em: Brazilian general's speech | Oversight Board. Acesso em 16 fev. 2025).

[353] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 243.

e indivíduos podem utilizar para moldar sua compreensão e participação no discurso público, nas eleições e em outros aspectos fundamentais da democracia:[354]

> Para muitos defensores da transparência de plataformas, o objetivo não é apenas proteger os direitos dos indivíduos como consumidores, mas capacitá-los como participantes do discurso público e da autogovernança democrática.
> Esse é o argumento da transparência que, pessoalmente, considero mais convincente. Devemos esperar e exigir mais transparência de plataformas como o *YouTube* e o *Facebook*, não apenas porque somos consumidores de seus produtos, mas também devido ao importante papel que as plataformas desempenham na formação de nosso ecossistema de informações e nos resultados políticos. Sem melhores informações sobre o papel que as plataformas desempenham, somos individual e coletivamente prejudicados no projeto de autogovernança democrática. Chamarei isso de 'interesse pela democracia'.[355]

É claro que saber sobre campanhas de desinformação nas redes sociais para manipulação de processos eleitorais, remoção excessiva de conteúdo para evitar responsabilização e/ou vieses discriminatórios de modelos automatizados podem ajudar na percepção do que as redes sociais pretendem proporcionar ao consumidor. Contudo, é particularmente mais relevante a exata compreensão do que essas distorções podem causar sobre a própria democracia.[356]

A garantia de transparência também permite a observância de outros direitos procedimentais, como o devido processo, que são reconhecidos e consagrados, ao lado de medidas de transparência, em padrões de liberdades individuais, princípios de moderação de conteúdo e princípios internacionais de direitos humanos, instrumentalizados em documentos de *softlaw*, como os Princípios de Manila e os Princípios de Santa Clara.[357-358]

---

[354] KELLER, Daphne. *Platform Transparency and the First Amendment…*, p. 31.
[355] *Ibidem*, p. 45: "For many platform transparency advocates, the goal is not merely to protect individuals' rights as consumers, but to empower them as participants in public discourse and democratic self governance. This is the transparency argument that I personally find most compelling. We should expect and demand better transparency from platforms like YouTube and Facebook, not just because we are consumers of their products, but also because of the major role platforms play in shaping our information ecosystem and political outcomes. Without better information about the role platforms play, we are individually and collectively impaired in the project of democratic self-governance. I will call this a 'democracy interest'."
[356] *Ibidem*, p. 46.
[357] MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*…, p. 7.
[358] Os Princípios de Manila e os Princípios de Santa Clara serão analisados mais detalhadamente no capitulo 3.3.3 *infra*, no qual se examinará as principias propostas de organizações nacionais e internacionais sobre a moderação por normas de procedimento.

Nos Princípios de Manila, grupos da sociedade civil de todo o mundo recomendam que "transparência e prestação de contas devem ser integradas em leis e em políticas e práticas de restrições de conteúdo" (sexto princípio), bem como que as "leis, políticas e práticas de restrição de conteúdo devem respeitar o devido processo" (quinto princípio) e as "requisições de restrição de conteúdos devem ser claras, não ambíguas e seguir o devido processo" (terceiro princípio), incluindo, por exemplo, o direito do usuário a ser ouvido antes da restrição de qualquer conteúdo com base em uma ordem ou requisição.[359]

Os Princípios de Santa Clara, organizado por um grupo de organizações de direitos humanos, defensores e acadêmicos, refletem recomendações semelhantes sobre a melhor forma de obter transparência e *accountability* na moderação de conteúdo feita pelas plataformas digitais. Esses princípios são organizados em três categorias: *(i)* princípios fundamentais, *(ii)* operacionais; e *(iii)* voltados aos governos e outros atores governamentais.[360]

Dentre os princípios fundamentais, destaca-se o dever das plataformas de "garantir que considerações sobre direitos humanos e devido processo sejam integradas em toda etapa do processo de moderação de conteúdos e divulgar informação mostrando como tal integração ocorre". Os princípios operacionais, por outro lado, buscam "divulgar informações sobre peças de conteúdo e contas acionados, analisado por pais ou região, se disponível, e a categoria de regra violada", para garantir "transparência" à moderação de conteúdo. E, por fim, em relação aos princípios voltados ao governo e outros atores governamentais, atribui-se a esses agentes o papel de "remover barreiras à transparência que previnam as empresas de cumprir plenamente os princípios supracitados". Nesses termos, a transparência das empresas "é um elemento crucial para assegurar confiança nos processos de moderação de conteúdo".[361]

Impulsionadas por instrumentos de *softlaw*, organizados pela sociedade civil e especialistas para demandar maior transparência e *accountability* das redes sociais, as maiores plataformas digitais passaram de forma voluntária a adotar a

---

[359] ELETRIC FRONTIER FOUNDATION *et al*. *Princípios de Manila sobre Responsabilidade de Provedores*. Disponível em: <https://manilaprinciples.org/index.html>. Acesso em 15 dez. 2024.
[360] ACESS NOW *et al*. *Os Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo*. Disponível em: <https://santaclaraprinciples.org/>. Acesso em 12 jan. 2024.
[361] *Idem*.

transparência como um mecanismo de governança para aumentar a confiança do público na operação de seus sistemas de moderação de conteúdo.[362]

Atualmente, as grandes plataformas divulgam relatórios de transparência, independentemente de qualquer obrigação legal ou delimitação das informações necessárias. Na ausência de regulação específica, as plataformas decidem quais informações serão apresentadas ou não ao público. Embora essa prática represente um avanço em termos de transparência, esses relatórios apresentam lacunas significativas e não disponibilizam informações padronizadas entre si, o que limita a realização de análises comparativas e impede a adequada compreensão sobre o ambiente digital e as práticas de moderação adotadas por essas grandes plataformas.

2.3.1.2. Relatórios de transparência

As grandes plataformas digitais, *Facebook*, *X* e *Youtube*, começaram a publicar em 2018 os primeiros relatórios de transparência sobre a aplicação das diretrizes de comunidade (*Community Guidelines Enforcement Reports*). Esses relatórios apresentam limitações significativas. Os dados divulgados nos relatórios de transparência refletem a avaliação das plataformas, sem fornecer informações dos casos subjacentes. Isso impede uma avaliação externa de pesquisadores a respeito da precisão das decisões de moderação ou identificação de tendências inconsistentes na aplicação das regras.[363]

A maioria dos relatórios de transparência abrange categorias específicas de remoções, frequentemente limitadas a ações governamentais ou de detentores de direitos autorais. Os relatórios de transparência também variam quanto aos detalhes das informações e classificações dos dados divulgados, o que dificulta a realização de análises comparativas entre as redes sociais. Um relatório que contabiliza quantas notificações foram recebidas pela plataforma não pode ser comparado, por exemplo, com outro que registra a quantidade de solicitações de remoção, já que em uma notificação podem ser formuladas mais de uma solicitação de retirada de conteúdo.[364]

---

[362] MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*…, p. 7.
[363] KELLER, Daphne; LEERSSEN, Paddy. *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* …, p. 15.
[364] *Idem*.

O exemplo dos relatórios das grandes plataformas ilustra a variedade de detalhes das informações divulgadas por essas empresas. O relatório do *Youtube* indica o número de canais e vídeos removidos com base em 11 tipos distintos de violações, como *spam*, nudez e promoção de violência e extremismo, além de informar como as violações foram identificadas. O Facebook disponibiliza informações mais abrangentes, detalhando a frequência com que usuários apelaram contra decisões de remoção e as ocasiões em que o conteúdo foi restaurado posteriormente, seja de forma proativa ou por meio de recursos dos usuários, bem como dados de suas operações, como o quantitativo de sua equipe de moderação de conteúdo, composto por mais de 20.000 moderadores humanos. Além disso, o Facebook divulgou uma versão pública acessível de seus Padrões da Comunidade e um guia para ajudar a interpretar os números do seu relatório. O relatório do *X*, por sua vez, é significativamente menos detalhado, apresentando apenas o número de contas denunciadas e as ações tomadas em seis categorias de violações, sem fornecer informações sobre recursos, taxas de reintegração ou mecanismos de denúncia, exceto aqueles de entidades governamentais reconhecidas.[365]

Mais recentemente, essas grandes plataformas passaram a publicar relatórios específicos como parte de sua conformidade com a regulação alemã (*NetzDG*), conhecida por suas rígidas regras de remoção de conteúdo e pela imposição de diversos requisitos a serem preenchidos em relatórios semestrais de transparência, que envolvem tanto dados estatísticos quanto detalhes operacionais.

Na segunda metade de 2018, o *YouTube* recebeu notificações, com base na *NetzDG*, que identificavam mais de 250.000 itens, com preponderância para discursos de ódio e extremismo político, seguida de difamação e insultos e, ainda, de conteúdo sexual. Em resposta, o *YouTube* removeu 54.644 itens, com taxas de remoção variando conforme a categoria de conteúdo. O relatório semestral também indica se as notificações foram enviadas por usuários ou por agências governamentais alemãs. A análise do relatório demonstra, ainda, que o *YouTube* baseou-se não apenas nas suas "Diretrizes da Comunidade", mas também na própria legislação alemã para definir o *status* legal de mais de 10.000 itens.[366]

Os relatórios da *NetzDG* do Facebook apresentam uma outra imagem. Nesse período, a plataforma reportou ter recebido apenas 500 reclamações com base na

[365] *Ibidem*, p. 16.
[366] *Ibidem*, p. 23.

NetzDG, envolvendo tão somente 1.048 itens de conteúdo – uma fração muito menor do que o *YouTube*. No entanto, as remoções baseadas nos seus "Padrões de Comunidade" nesse período somaram milhões, e não foram incluídas no relatório da *NetzDG*. A disponibilização de relatório incompleto, criando uma imagem distorcida de sua moderação de conteúdo, deu ensejo à aplicação de multa ao *Facebook*, em 2 de julho de 2019, no valor de 2 milhões de euros, pelo Escritório Federal de Justiça da Alemanha. [367]

Uma importante avaliação externa sobre o grau de transparência das grandes plataformas digitais foi conduzida por pesquisadores da *Electronic Frontier Foundation's*, uma organização sem fins lucrativos sediada em San Francisco, na Califórnia/EUA, cujo objetivo declarado é proteger os direitos de liberdade de expressão na era digital, através do relatório *Who Has Your Back? Censorship Edition 2019*.[368]

Nesse relatório, foram examinadas as políticas de moderação de conteúdo das principais redes sociais, em diversas categorias que envolvem *(i)* transparência nas solicitações de governos com base na legislação local; *(ii)* transparência nas solicitações de governos com base em violações às políticas das plataformas; *(iii)* transparência nas notificações aos usuários sobre cada remoção de conteúdo e suspensão de conta; *(iv)* a disponibilização aos usuários de um processo de apelação para contestar remoções de conteúdo e suspensões de contas; *(v)* transparência com relação ao número de apelações; e *(vi)* apoio público aos Princípios de Santa Clara.[369]

---

[367] *Ibidem*, p. 24.

[368] CROKER, Andrew; GEBHART, Gennie; MACKEY, Aaron; OPSAHL, Kurt; TSUKAYAMA, Hayley; WILLIAMS, Jamie Lee; YORK, Jillian C. *Who Has Your Back? Censorship Edition 2019*. Disponível em: <https://www.eff.org/files/2019/06/11/whyb_2019_report.pdf>. Acesso em 20 jan. 2025.

[369] *Ibidem*, p. 6: "This year's Who Has Your Back report examines major tech companies' content moderation policies in the midst of massive government pressure to censor. We assess companies' policies in six categories: ● Transparency in reporting government takedown requests based on legal requests ● Transparency in reporting government takedown requests alleging platform policy violations ● Providing meaningful notice to users of every content takedown and account suspension ● Providing users with an appeals process to dispute takedowns and suspensions ● Transparency regarding the number of appeals ● Public support of the Santa Clara Principles." Tradução: "O relatório Who Has Your Back deste ano examina as políticas de moderação de conteúdo das principais empresas de tecnologia das principais empresas de tecnologia em meio à pressão maciça do governo para censurar. Avaliamos as políticas das empresas em seis categorias: ● Transparência na comunicação de solicitações de remoção do governo com base na legislação local; ● Transparência na comunicação de solicitações de remoção do governo com base em violações à política da plataforma; ● Envio de notificações aos usuários sobre cada remoção de conteúdo e

O Relatório conclui que nenhuma das três plataformas digitais, *Facebook*, *Youtube* e *X*, atinge a pontuação máxima de grau de transparência definido com base nessas seis categorias pré-definidas. Para ilustrar a pontuação de cada uma dessas redes sociais de forma individualizada, confira-se os recortes extraídos do quadro geral elaborado pelos pesquisadores da *Electronic Frontier Foundation's*, no relatório *Who Has Your Back? Censorship Edition 2019*:

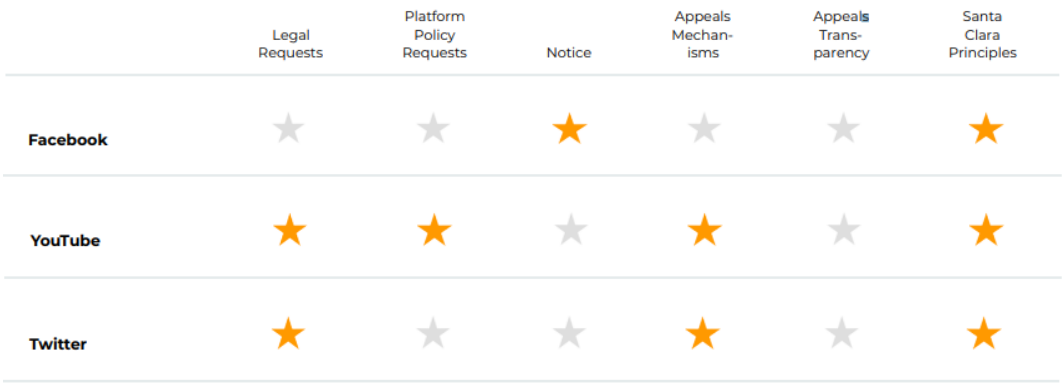

Figura 2 – Pontuação de grau de transparência das três principais redes sociais no relatório da Who Has Your Back? Censorship Edition 2019

O *Facebook* somente atingiu dois pontos nas seis categoriais. Isso porque, a plataforma não reporta em seus relatórios de transparência o número de solicitações governamentais recebidas seja com base na legislação local ou em violação às políticas da plataforma e não possibilita apelações contra todas as penalidades baseadas em violações de seus Padrões de Comunidade. O Facebook apenas pontua por promover notificações para remoções de conteúdo e violações às Padrões de Comunidade e apoiar publicamente os Princípios de Santa Clara.[370]

O *Youtube*, por sua vez, pontua em quatro das seis categorias. Essa pontuação se justifica pelo reporte de solicitações governamentais recebidas seja com base na legislação local ou em violação às políticas da plataforma, pela possibilidade de os usuários apresentarem apelações contra decisões de remoção e suspensão e, ainda, pelo apoio público aos Princípios de Santa Clara. A plataforma deixa de marcar pontos por não se comprometer publicamente a notificar os

suspensão de conta suspensão de contas; ● Disponibilização aos usuários de um processo de apelação para contestar remoções e suspensões; ● Transparência com relação ao número de apelações; e ● Apoio público aos Princípios de Santa Clara."

[370] *Ibidem*, pp. 17-18.

usuários sobre remoções legais, bem como não informar número ou o resultado dos recursos apresentados por usuários.[371]

Já o *X*, antigo *Twitter*, pontua em três categorias, por disponibilizar seção específica sobre pedidos de remoção que incluem solicitações governamentais com base na legislação local, permitir que usuários apelem contra decisões de remoção de conteúdo e suspensão de contas e, ainda, apoiar publicamente os Princípios de Santa Clara. A plataforma deixa de marcar pontos nas outras três categorias, porque não reporta solicitações governamentais fundadas em violação às políticas da plataforma, deixa de notificar os usuários em caso de remoção legal relacionada a atos de "terrorismo" e não fornece o número total ou o resultado dos recursos apresentados pelos usuários.[372]

Além desse ilustrativo exemplo, outras valiosas análises sobre as falhas no cumprimento dos desafios de transparência e *accountability* foram conduzidas em relatório elaborado pela *New America's Open Technology Institute* (OTI) e pela *Harvard University's Berkman Center for Internet & Society*, intitulado *The Transparency Reporting Toolkit*, que compara os relatórios de transparência disponibilizados pelas principais plataformas e as recomendações de melhores práticas de transparência, como, por exemplo, os Princípios de Santa Clara.[373]

As falhas cometidas pelas plataformas digitais no cumprimento dos deveres de transparência e *accountability* causam prejuízos aos usuários, na medida em que ameaçam sua capacidade de compreendem os limites impostos à sua liberdade de expressão e de buscarem remédios adequados quando seus direitos são violados, conforme apontam as pesquisas conduzidas por David Kaye, relator especial das Nações Unidas para Liberdade de Opinião e Expressão entre 2014 e 2020.[374]

Além disso, a ausência de dados completos sobre as regras e os sistemas de moderação de conteúdo impede que seja identificada a origem e o alcance da desinformação, bem como que seja avaliada a eficiência das medidas adotadas e os impactos dos algorítmicos sobre elas. Essas foram conclusões de um trabalho

---

[371] *Ibidem*, pp. 39-40.
[372] *Ibidem*, pp. 35-36.
[373] BUDISH, Ryan; WOOLERY, Liz; e BANKSTON, Kevin. The Transparency Reporting Toolkit: Survey & Best Practice Memos for Reporting on U.S. Government Requests for User Information. *New America's Open Technology Institute (OTI) and Harvard University's Berkman Center for Internet & Society*. Disponível em: <https://www.newamerica.org/oti/policy-papers/the-transparency-reporting-toolkit/>. Acesso em 20 jan. 2025.
[374] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 242.

publicado, em 2020, pela *Harvard Kennedy School Misinformation Review*, que reforça a importância do compartilhamento de dados ao público e pesquisadores, trazendo a perspectiva de que esses dados poderiam constituir "um bem cultural comum", que podem viabilizar "pesquisas vitais sem o envolvimento ou o controle de interesses corporativos".[375]

Embora as plataformas digitais tenham adotado voluntariamente a transparência dentre os seus mecanismos de governança, visando o aumento da confiança pública em seus sistemas de moderação de conteúdo, os dados empíricos analisados em estudos realizados de forma independente apontam para falhas sistemáticas de transparência e *accountability* das redes sociais. Essas falhas podem

[375] PASQUETTO, Irene V, et. al. Tackling misinformation: what researchers could do with social media data. *Harvard Kennedy School (HKS) Misinformation Review*, 2020. Disponível em: <https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/>. Acesso em 21 jan. 2025: "The information ecology produced by social media platforms and the data they collect is part of our cultural heritage. These data represent "libraries" of the present and the future, and they should be considered cultural artefacts that, like their physical counterparts, ought to be "owned" by the people who produced them—which is all of us. We must reclaim the information ecology for the people who created it. This is crucial for independent research in the public interest: to conduct such "dream research," researchers must not be supplicants to social-media companies but should go straight to the users, recruiting people as "citizen scientists" and knowledge-co-creators. This can be achieved via dedicated research platforms and browser plug-ins, mobile applications, and other digital data collection tools that allow researchers access to people's online activity subject to strict confidentiality and anonymity constraints. These data could constitute a common cultural good, permitting vital research without involvement or control by corporate interests. Some of the questions we could then address include examinations of (a) how malicious disinformation (e.g., concerning COVID-19) affects subsequent content engagement, and how this is shaped by countermeasures already in place or being designed by empirical research; (b) people's motives to share information, in particular intentional sharing of information that is known to be false; (c) drivers of polarization and de-polarization; and (d) ways in which a common ground for evidence and rules of arguments can be re-established." Tradução: "A ecologia de informações produzida pelas plataformas de mídia social e os dados que elas coletam fazem parte do nosso patrimônio cultural. Esses dados representam "bibliotecas" do presente e do futuro, e devem ser considerados artefatos culturais que, assim como suas contrapartes físicas, devem ser "propriedade" das pessoas que os produziram - que somos todos nós. Devemos reivindicar a ecologia da informação para as pessoas que a criaram. Isso é fundamental para a pesquisa independente de interesse público: para realizar essa "pesquisa dos sonhos", os pesquisadores não devem ser suplicantes das empresas de mídia social, mas devem ir diretamente aos usuários, recrutando pessoas como "cientistas cidadãos" e cocriadores de conhecimento. Isso pode ser feito por meio de plataformas de pesquisa dedicadas e plug-ins de navegador, aplicativos móveis e outras ferramentas de coleta de dados digitais que permitem que os pesquisadores acessem a atividade on-line das pessoas, sujeita a restrições estritas de confidencialidade e anonimato. Esses dados poderiam constituir um bem cultural comum, permitindo pesquisas vitais sem o envolvimento ou o controle de interesses corporativos. Algumas das questões que poderíamos abordar incluem exames de (a) como a desinformação maliciosa (por exemplo, em relação à COVID-19) afeta o engajamento de conteúdo subsequente e como isso é moldado por contramedidas já em vigor ou que estão sendo projetadas por pesquisas empíricas; (b) os motivos das pessoas para compartilhar informações, em particular o compartilhamento intencional de informações que são sabidamente falsas; (c) os fatores de polarização e despolarização; e (d) as maneiras pelas quais um terreno comum para evidências e regras de argumentos pode ser restabelecido."

e devem ser objeto de uma intervenção regulatória pública, conforme será abordado adiante no capítulo 3 *infra*.

3. AUTORREGULAÇÃO REGULADA: A REGULAÇÃO BIFÁSICA DA LIBERDADE DE EXPRESSÃO NAS REDES SOCIAIS

Conforme demonstrado nos capítulos anteriores, a moderação de conteúdo realizada pelas redes sociais com base em seus termos de uso não configura violação à liberdade de expressão, tampouco consiste em censura prévia. O modelo de autorregulação das redes sociais, que embasa a regulação privada do discurso público na era digital, está fundado no exercício da autonomia privada e da liberdade editorial das plataformas digitais.

Nesse contexto, as ações realizadas na aplicação das regras e políticas internas de moderação de conteúdo pelas plataformas, previamente estabelecidas e aceitas pelos usuários, podem ser classificadas como sanções contratuais. A execução dessas sanções não depende da intervenção judicial, podendo ser aplicadas forçadamente por meio dos próprios sistemas informatizados das redes sociais.

O exercício desse direito das plataformas, por meio de suas práticas de moderação de conteúdo com base nos seus termos de uso, está sujeito a limites negativos que decorrem dos direitos dos usuários, como, por exemplo, os direitos à liberdade de expressão, privacidade, honra, segurança, proteção de dados etc. A eficácia dos direitos dos usuários na relação com as plataformas impõe também o cumprimento deveres específicos procedimentais de transparência e *accountability*. Esses deveres representam, como visto acima, verdadeiros desafios para as redes sociais, mas são fundamentais para garantir a legitimidade decisória das plataformas digitais e para promover valores democráticos no ambiente virtual.

Diante das falhas sistemáticas das plataformas digitais no cumprimento dos deveres e desafios de transparência, *accountability* e legitimidade decisória, é razoável vislumbrar a intervenção regulatória pública (limite positivo), para resguardar interesses públicos relevantes, dentre os quais, destaca-se a própria liberdade de expressão dos usuários.

Neste capítulo, será abordada a proposta normativa bifásica da corregulação ou autorregulação regulada, para endereçar os problemas oriundos da regulação privada do discurso público nas redes sociais. Para tanto, será preciso analisar,

inicialmente, os limites e riscos decorrentes da heterorregulação ou regulação puramente pública, que inviabilizam essa solução regulatória.

Em uma segunda etapa, será analisada uma possível resposta regulatória bifásica, através da autorregulação regulada ou corregulação, com a participação ativa dos agentes estatais e dos agentes regulados (modelo policêntrico), que se mostre adequada às complexidades inerentes às redes sociais. Nesse contexto, será proposto o estabelecimento de deveres procedimentos de transparência, devido processo e isonomia para a moderação de conteúdo, cujo descumprimento sistêmico poderia ensejar a responsabilização das plataformas.

Ao final deste capítulo, serão analisadas as principais propostas regulatórias e de boas práticas de conceituadas organizações nacionais e internacionais a respeito da moderação de conteúdo, bem como o papel de um possível órgão fiscalizador independente no contexto regulatório proposto.

3.1. O Estado não deve deter o monopólio da regulação das redes sociais: riscos da heterorregulação ou da regulação puramente pública

A intervenção do Estado sobre o exercício do direito à liberdade de expressão sempre foi vista com grande desconfiança e resistência, diante do caráter autoritário e antidemocrático a que historicamente esteve atrelado. Essa percepção não é diferente em relação à regulação das redes sociais. A preocupação em torno dos riscos de o Estado agir de acordo com interesses próprios, para censurar discursos que considere prejudiciais, é equivalente ou, até mesmo, superior à regulação privada exercida pelas plataformas digitais.[376]

Uma intervenção estatal indevida sobre a liberdade de expressão dos usuários pode ocorrer, por exemplo, por meio de *(i)* leis excessivamente rígidas de responsabilização civil das plataformas digitais por conteúdo publicado por terceiro, que criam incentivos para a remoção excessiva; *(ii)* leis vagas que utilizem conceitos indeterminados para, em nome da "paz social" ou do "interesse público", impor restrições à liberdade de expressão; e *(iii)* responsabilização civil e criminal de representantes legais localizados no país em caso de não atendimento a ordens estatais de remoção de conteúdo. Além disso, os Estados podem tentar restringir

[376] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 220.

discursos por meio extralegais, como ameaças de regulação excessiva e denúncias de conteúdo pelos meios disponibilizados pelas próprias plataformas.[377]

Essas medidas relevam a possibilidade de haver um interesse do Estado em censurar discursos por meio das próprias empresas de tecnologia. Existem, ao menos, quatro motivos que podem levar ao interesse estatal em efetivar a censura através das plataformas: *(i)* a capacidade técnica de moderação e controle de discurso das plataformas digitais é mais amplo, porque elas controlam os códigos e os algoritmos que definem a escala de disseminação de conteúdo; *(ii)* as plataformas digitais são a fonte primária de coleta e processamento de dados dos usuários, que permite identificar o que esses usuários estão fazendo e publicando nas redes sociais, facilitando o controle de discursos; *(iii)* é menos custosa a regulação de poucas plataformas digitais que podem facilmente regular o discurso de milhões de pessoas do que regular os cidadãos individualmente; e, por fim, *(iv)* esse tipo de censura evita maior exposição do governo, por ser mais obscura e aparentemente sem relação direta com a atividade estatal.[378]

A possibilidade de efetivação da censura estatal por intermédio das empresas de tecnologia, viabilizando o controle de discursos em larga escala, traz consigo os já mencionados riscos de *censura colateral*, que decorrem da remoção excessiva de conteúdos lícitos pelas redes sociais a fim de evitar sanções ou ameaças regulatórias, bem como da *censura prévia*, relacionados às exigências governamentais de filtragem e moderação de conteúdo feitas às plataformas.[379]

Os riscos da regulação puramente estatal podem ser potencializados por meio de iniciativas legislativas que busquem definir o conteúdo daquilo que *pode* ou *não pode* ser publicado nas redes sociais, em virtude da dificuldade de se estabelecer "um consenso na definição dos limites da liberdade de expressão". É comum que haja razoável desacordo em relação à proteção constitucional de determinadas manifestações, que margeiam temas complexos e com potenciais riscos de configurar discurso de ódio.[380-381]

---

[377] *Idem*.
[378] *Ibidem*, p. 99.
[379] *Ibidem*, pp. 100-101.
[380] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 221.
[381] Um exemplo de medidas de moderação que causaram controvérsia foi o bloqueio, em janeiro de 2021, das contas de Donald Trump do *X*, *Facebook* e *Instagram*. (PORTAL G1. *Twitter, Facebook e Instagram bloqueiam contas de Trump temporariamente*. Publicado em 6 jan. 2025. Disponível

A regulação puramente estatal também encontra, no entanto, limites negativos nos direitos dos usuários e das plataformas digitais. A intervenção regulatória do Estado não pode "silenciar" os usuários, promovendo censura prévia ou colateral, em violação ao direito fundamental à liberdade de expressão. Da mesma forma, a heterorregulação não pode impedir que as plataformas digitais tenham autonomia para criar o tipo de comunidade *online*, desde que cumpridos os deveres de transparência e *accountability*, sob pena de violar o direito à livre iniciativa e à liberdade de expressão das próprias redes sociais.[382]

No contexto brasileiro, a liberdade de expressão ou a garantia contra a censura estatal deve considerar a ideia da "cultura do silêncio", conceito construído por Paulo Freire.[383-384] A "cultura do silêncio" remete à noção de que a sociedade brasileira se nega à comunicação e ao diálogo e, em seu lugar, busca oferecer apenas "comunicados"; um ambiente de tolhimento da voz e de ausência de comunicação. A análise de Freire tem como ponto de partida os seus estudos sobre a herança do período colonial, a estrutura de dominação criada nesse período e sua influência sobre as representações e os comportamentos dos brasileiros.[385]

É nesse contexto que se vislumbra os riscos de surgirem iniciativas governamentais que visem o sufocamento das vozes dos brasileiros nas redes sociais. Alguns exemplos são ilustrativos da existência desses riscos. Em 2016, o centro de pesquisas InternetLab divulgou uma pesquisa com dois dados

---

em: <https://g1.globo.com/economia/tecnologia/noticia/2021/01/06/twitter-diz-que-conta-de-trump-ficara-bloqueada-por-12-horas.ghtml>. Acesso em 9 fev. 2025).

[382] A respeito desses limites negativos, Domingos Soares Farinho afirma: "[a] fixação de tarefas administrativas respeitantes ao funcionamento das redes sociais nunca poderá implicar uma intervenção pública, seja ela legislativa ou administrativa, que comprima de modo ilegal (inconstitucional) o conteúdo de direitos e liberdades dos titulares das redes sociais e dos seus utilizadores. Mas deve assegurar que o exercício da auto-ordenação ou autorregulação não prejudica o normal exercício de liberdades fundamentais dos utilizadores." (FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., p. 60).

[383] FREIRE, Paulo. *Pedagogia do oprimido*, 17ª. Ed. Rio de Janeiro, Paz e Terra, 1987, p. 26. Disponível: <https://pibid.unespar.edu.br/noticias/paulo-freire-1970-pedagogia-do-oprimido.pdf/view>. Acesso em 26 jan. 2025: "Acontece, porém, que, ao passarem de exploradores ou de expectadores indiferentes ou de herdeiros da exploração – o que é uma conivência com ela – ao pólo dos explorados, quase sempre levam consigo, condicionados pela 'cultura do silêncio', toda a marca de sua origem. Seus preconceitos. Suas deformações, entre estas, a desconfiança do povo. Desconfiança de que o povo seja capaz de pensar certo. De querer. De saber."

[384] VALENTE, Mariana Giogetti. A liberdade de expressão na internet: Da utopia à era das plataformas. In: FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*. São Paulo: Perspectiva, 2020, p. 31.

[385] DE LIMA, Venício A. Da cultura do silêncio ao direito à comunicação. *Observatório da Impresa*. 22 nov. 2011. Disponível: <https://www.observatoriodaimprensa.com.br/feitos-desfeitas/da-cultura-do-silencio-ao-direito-a-comunicacao/>. Acesso em 26 jan. 2025.

interessantes sobre a influência de agentes políticos sobre decisões do Judiciário: *(i)* um terço dos processos judiciais ajuizados para remoção de conteúdos de humor na internet são movidos por políticos; e *(ii)* esses processos têm uma alta taxa de êxito nos tribunais brasileiros, com a concessão de indenização pleiteada em mais da metade dos processos julgados em segunda instância.[386]

No ano de 2017, em meio à discussão sobre a reforma política, o Congresso Nacional inseriu no texto aprovado, na madrugada do dia 5 de outubro, uma medida obrigando as plataformas digitais a remover informações falsas e ofensas em desfavor de partido, coligação ou candidato, sem necessidade de prévia ordem judicial, sob pena de responsabilização. A medida foi vetada pelo então presidente Michel Temer, em virtude da forte reação de entidades de imprensa e de organizações da sociedade civil que integram a Coalizão Direitos nas Redes.[387]

Além disso, inúmeros projetos de lei foram apresentados nos últimos anos perante o Congresso Nacional com medidas de responsabilização civil imediata das plataformas digitais e de criminalização de cidadãos que se manifestem ou repassem "notícias falsas".[388]

O cenário brasileiro releva uma alta conflituosidade social em torno da política institucional sobre a regulação das redes sociais, o que agrava o risco e a preocupação de o governo recorrer a processos judiciais e iniciativas legislativas para censurar opiniões dissonantes. Isso não significa, contudo, que se nega ao Estado o papel de promoção da liberdade de expressão, inclusive por meio da intervenção regulatória. Esse papel é fundamental para os casos em que o próprio discurso *online* pode ter um efeito silenciador, como é o caso do discurso de ódio.[389]

Portanto, considerando os riscos de uma intervenção estatal autoritária ou excessiva, com efeito silenciador aos usuários e violador dos direitos das plataformas digitais, é que se defende que o Estado não deve deter o monopólio da regulação das redes sociais. No entanto, como não se pode ignorar o papel do Estado na construção de uma solução normativa, é preciso discutir um modelo de regulação que envolva não apenas o Estado, mas também outros agentes que integram o

---

[386] VALENTE, Mariana Giogetti. A liberdade de expressão na internet: Da utopia à era das plataformas. In: FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*..., p. 31.
[387] *Ibidem*, p. 32.
[388] Por exemplo, os projetos de lei sob os nºs. 9.532/2018, 9.973/2018, 9.838/2018 e 8.043/2017, na Câmara dos Deputados; e 218/2018 e 473/2017, no Senado Federal (*Ibidem*, p. 31).
[389] *Ibidem*, p. 32.

ecossistema das redes sociais, em especial as próprias plataformas digitais, para que se promova a liberdade de expressão, e não a censura de discursos lícitos.

3.2. Modelo policêntrico: cooperação entre Estado regulador e atores ou setores sociais a serem regulados

Enquanto os modelos de autorregulação e de regulação puramente estatal parecem insuficientes para endereçar as problemáticas no ambiente *online*, a autorregulação regulada ou corregulação emerge como uma alternativa promissora, capaz, inclusive, de desenvolver mecanismos de legitimação das decisões de moderação das plataformas relacionadas ao exercício da liberdade de expressão.[390]

O marco distintivo entre a autorregulação regulada em relação aos demais modelos de regulação acima tratados é que nela se recorre à cooperação entre os agentes estatais e os agentes privados a serem regulados. Esse modelo contribui para o uso da expertise e conhecimento técnico das plataformas digitais na garantia e proteção de direitos fundamentais e valores de interesse público. [391]

Abre-se, então, o caminho para a construção de um modelo que a doutrina tem chamado de "policêntrico", que não está restrito a uma única fonte regulatória, mas sim à conjunção entre fontes privadas e públicas, para a regulação do ambiente digital.[392] Trata-se, portanto, de um modelo regulatório focado na cooperação entre o Estado regulador e os atores ou setores sociais a serem regulados.[393]

Essa é, aliás, uma imposição do próprio sistema jurídico no contexto do mundo digital, que envolve direitos fundamentais tanto das plataformas digitais, que precisam ser assegurados – como, por exemplo, estipular, implementar e fazer cumprir as regras privadas de utilização de seus sistemas de moderação –, quanto dos seus usuários, que devem ser tutelados pelo Estado regulador em face das possíveis violações aos seus direitos decorrentes da regulação privada do discurso público nas redes sociais.

---

[390] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 221.
[391] *Ibidem*, p. 223.
[392] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., pp. 43-104, p. 58.
[393] ABBOUD, Georges; CAMPOS, Ricardo. A autorregulação regulada como modelo do Direito procedualizado: regulação de redes sociais e proceduralização. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, pp. 135-162, p. 144.

A cooperação entre agentes privados e públicos é fundamental para a regulação. Afinal, um dos problemas para o Estado intervir no mundo digital é a sua falta de conhecimento para promover uma intervenção eficiente em ambientes digitais dinâmicos, como as redes sociais, que envolvem expressivos e constantes avanços tecnológicos. Esse problema se agrava com a tendência cada vez mais forte de utilização de mecanismos automatizados para a moderação de conteúdo, por meio de ferramentas de inteligência artificial.[394]

Com efeito, a incerteza e a celeridade das transformações no mundo digital demandam maior criatividade e experimentalismo para se debater, cogitar e propor uma regulação que possa lidar, de forma eficiente, com essa árdua tarefa de acompanhar o caráter dinâmico das redes sociais.[395] Nesse sentido, Juliano Maranhão e Ricardo Campos afirmam:

> A autorregulação regulada oferece outra e nova possibilidade de lidar com as incertezas, ao conciliar vantagens das duas abordagens alternativas: a regulação *versus* a autorregulação. Foca-se no importante momento de auto-organização conforme expertise e dinâmica própria da indústria, estimulando-se, porém, alguns parâmetros gerais de interesses públicos caros ao Estado e sociedade. Nesse sentido, a autorregulação regulada consegue "induzir" o setor privado a contribuir para o cumprimento de tarefas públicas. Essa forma de regulação pode lidar melhor com uma sociedade que cada vez mais se locomove, e se distancia, de uma sociedade centrada em organizações conseguindo absorver melhor as incertezas e construir parâmetros melhores de eficácia na regulação.[396]

No modelo da autorregulação regulada ou corregulação, as plataformas digitais poderão, portanto, formular, interpretar e implementar a regulação, reduzindo os custos públicos e trazendo mais eficiência à implementação das normas regulatórios, por deterem o conhecimento técnico especializado sobre o funcionamento desses sistemas. Por sua vez, os agentes estatais ficarão responsáveis pelo estabelecimento de parâmetros de interesse público.[397]

---

[394] *Ibidem*, p. 138.
[395] MARANHÃO, Juliano; CAMPOS, Ricardo. Fake News e autorregulação regulada das redes sociais no Brasil: fundamentos constitucionais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., p. 343.
[396] *Ibidem*, p. 344.
[397] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 223.

Por estar fundada em um mecanismo de cooperação e coparticipação entre os agentes estatais e privados a serem regulados, é possível considerar o modelo de autorregulação regulada ou corregulação como uma forma moderna de regulação que, ao contrário da heterorregulação e da autorregulação, atende às "condições de possibilidade de regulação de âmbitos complexos como do mundo digital".[398] O caráter inovador desse modelo de regulação está centralizado na ideia de "incorporar elementos da auto-organização do setor privado e, ao mesmo tempo, não abrir mão completamente da implementação ou estruturação de interesses públicos mesmo que por via indireta".[399]

A intervenção regulatória do Estado tem o papel de assegurar interesses públicos relevantes, como a privacidade, liberdade de expressão, segurança de menores de idade, proteção de direitos autorais, prevenção e repressão a crimes no ambiente das redes sociais.[400] Essa intervenção deve ocorrer por meio de abordagens regulatórias que causem o mínimo de impacto possível sobre as atividades econômicas dessas plataformas, cuja expertise e conhecimento sobre os seus sistemas (*i.e.*, as redes sociais) são essenciais para o adequado desenho desse modelo de regulação.

A atuação do Estado regulador serve também ao propósito de corrigir uma das distorções do sistema jurídico das plataformas digitais: a utilização do Poder Judiciário como regulador da autorregulação das plataformas digitais ao invés do papel de meta-ponderador ou ponderador da ponderação corregulatória.[401_402]

---

[398] ABBOUD, Georges; CAMPOS, Ricardo. *A autorregulação regulada como modelo do Direito procedualizado: regulação de redes sociais e proceduralização*..., p. 139.
[399] *Ibidem*, p. 140.
[400] FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*..., p. 95.
[401] *Ibidem*, p. 94.
[402] Nesse sentido, Domingos Soares Farinho pondera que "uma tal regulação judicial oferece várias dificuldades e denota a própria dificuldade na sua aceitação. Em primeiro lugar, os tribunais, ao contrário dos reguladores legislativos ou administrativos não dispõem de poder de iniciativa ou sequer podem procurar um ponto de encontro comum com as redes sociais para regular o quadro jurídico das relações que se estabelecem nas plataformas. Além disso, os tribunais não decidem de forma geral, podendo apenas fixar jurisprudência que influencie casos futuros. Os tribunais estão limitados pelo princípio do pedido e pela eficácia *inter partes* de suas decisões. Em terceiro lugar, os tribunais têm uma limitação técnica de jurisdição quanto ao que podem julgar e condenar quanto a uma rede social, uma vez que as medidas que estas possam desenvolver implicam a sua efetiva exequibilidade técnica, como no caso de um mecanismo de filtragem de publicações idênticas ou semelhantes a uma publicação no *Facebook* julgada ilegal. Assim, a intervenção judicial, sendo incontornável, estará limitada quanto à sua eficácia, como estarão quaisquer modelos que assentem apenas ou fundamentalmente numa decisão judicial, se não existir um esteio pré-ponderador, dir-se-á, uma coponderação prévia partilhada entre sujeitos privados – as redes sociais – e sujeitos públicos – um regulador administrativo plural." (*Ibidem*, pp. 96-97).

O modelo da autorregulação regulada se revela promissor, justamente, por ser capaz de conciliar os interesses envolvidos no ambiente digital, por meio de um modelo de cooperação entre o Estado regulador e os atores ou setores sociais a serem regulados. Assim, é possível promover o interesse público que justifica a intervenção regulatória do Estado, enquanto são assegurados os interesses privados fundados na autonomia da vontade e na liberdade editorial das plataformas digitais.

Para os fins desse trabalho de pesquisa, entende-se que o modelo de autorregulação regulada ou corregulação para as redes sociais deve focar no estabelecimento de um "paradigma de direito da proceduralização", como defendem Georges Abboud e Ricardo Campos.[403]

Conforme será abordado adiante (capítulo 3.3 *infra*), a autorregulação regulada deverá estabelecer normas procedimentais, que contribuam para criar um procedimento adequado, capaz de assegurar o cumprimento de deveres específicos de transparência, *accountability*, devido processo e isonomia na moderação de conteúdo realizada pelas plataformas digitais. Afinal, o cumprimento desses deveres procedimentais específicos é condição necessária para garantir a legitimidade decisória das redes sociais.

3.3. Autorregulação regulada: a regulação por normas procedimentais

Na nova era digital, as transformações tecnológicas criaram novos espaços de manifestação e decisão, onde as plataformas digitais privadas definem, com relativa independência, aquilo que pode ou não ser dito nas suas redes sociais,[404] desafiando as compreensões tradicionais sobre a liberdade de expressão.

Ao longo dos últimos anos, esse novo modelo decisório, caracterizado pela opacidade dos sistemas de moderação de conteúdo das plataformas digitais, tem dado lugar à busca por maior transparência e participação dos usuários e de setores da sociedade civil.[405] Apesar dos avanços em termos de transparência, conforme

[403] ABBOUD, Georges; CAMPOS, Ricardo. *A autorregulação regulada como modelo do Direito proceduralizado*..., p. 140.
[404] RAMOS, Carlos Eduardo Vieira. *O Direito das plataformas: procedimento, legitimidade e constitucionalização na regulação privada da liberdade de expressão na Internet*. Dissertação (Mestrado em Direito) – Faculdade de Direito, Universidade de São Paulo, 2020, p. 22.
[405] *Ibidem*, p. 392: "… a mudança na moderação de conteúdo se caracteriza como uma transição entre *opacidade* e *transparência* procedimentais. Concretamente, isso significou que, antes da

destacado acima (subcapítulo 2.3.2 *supra*), há muito ainda a ser feito para se assegurar o respeito às garantias fundamentais na moderação de conteúdo.

Dado às complexidades decorrentes da utilização de novas tecnologias, o conhecimento necessário para a tomada de decisões no ambiente virtual não se encontra exclusivamente no Estado, sendo necessária a criação de novas formas de geração de conhecimento dentro do direito regulatório estatal que incorpore o conhecimento advindo da sociedade sobre essas tecnologias.[406]

Esse novo paradigma aponta, como já visto acima, para "uma forma de regulação mais reflexiva, em que a observação e a incorporação de modelos de auto-organização da sociedade ganham preponderância."[407] Daí a solução da autorregulação regulada ou corregulação para regular a moderação de conteúdo nas redes sociais, a fim de tornar o processo decisório das plataformas na moderação de conteúdo mais transparente e participativo.

Por certo que diferentes arranjos regulatórios podem ser rotulados como autorregulação regulada ou corregulação, a proposta a ser apresentada neste trabalho busca garantir uma participação ativa tanto aos agentes regulados quanto aos agentes estatais, para a implementação de um modelo procedimental que seja sensível às circunstâncias técnicas do mercado digital e, ao mesmo tempo, promova princípios públicos como transparência, *accountability* e legitimidade.[408]

Noutras palavras, defende-se o estabelecimento de normas procedimentais, contando com a participação ativa de agentes estatais e regulados, em que "o conhecimento para decisão não se encontra na norma posta legitimamente pelo parlamento, nem em princípios abstratos, mas no procedimento estabelecido no direito posto."[409] A respeito do tema, Georges Abboud e Ricardo Campos acrescentam que:

---

mudança, não havia informações disponíveis sobre como as decisões sobre aquilo que as pessoas podem dizer eram tomadas: não se sabia, também, quando, como e se a plataforma havia agido sobre uma publicação; qual era o volume de sua atuação sobre aquilo que os usuários diziam. Também não eram claras as regras e a forma como o procedimento caminhava de um momento inicial, em que algo era denunciado aos moderadores – ou por eles detectado – até a decisão final, de manter ou não algo na plataforma."

[406] ABBOUD, Georges; CAMPOS, Ricardo. *A autorregulação regulada como modelo do Direito proceduralizado...*, p. 143.

[407] *Idem*.

[408] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo...*, p. 223.

[409] ABBOUD, Georges; CAMPOS, Ricardo. *A autorregulação regulada como modelo do Direito proceduralizado...*, p. 143.

> A mais-valia de um modelo de proceduralização do direito concretizado aqui no instituto da autorregulação regulada, como caso de regulação das redes sociais no tocante às notícias fraudulentas, decorrentes especialmente do fato de que, por um lado, ele não consiste numa forma de regulação mais direta, como das agências reguladoras, e, por outro lado, ele incorpora dentro de seu conceito regulatório a participação do setor privado objeto da regulação, incorporando um conhecimento de áreas tecnológicas, o qual o Estado não dispõe. Com isso, o modelo da autorregulação regulada responde ao *déficit* de conhecimento, gerando procedimentos (proceduralização) e uma abertura temporal do direito para lidar com uma sociedade cada vez mais complexa.[410]

Esse modelo de regulação visa a atingir também os seguintes objetivos: *(i)* reduzir a assimetria informacional entre usuários e plataforma; *(ii)* resguardar o direito à liberdade de expressão de intervenções excessivas; e *(iii)* proteger a democracia no ambiente digital. Mais especificamente, pretende-se, assim, *(i)* criar incentivos para que as plataformas digitais removam conteúdos ilícitos ou danosos, *(ii)* proteger conteúdos e discurso lícitos, evitando a remoção excessiva, e *(iii)* promover a liberdade de iniciativa e inovação.[411]

Esses objetivos estão alinhados, portanto, à implementação de um procedimento transparente, isonômico e garantidor das garantias fundamentais dos usuários das redes sociais, que permita aos usuários não apenas entender as regras de moderação de conteúdo, mas também participar do procedimento de moderação e, até mesmo, questionar as decisões tomadas pelas redes sociais.

A implementação de um procedimento adequado para a moderação de conteúdo das plataformas digitais contribuirá para a minimização de erros no processo decisório e para a legitimidade das decisões de moderação, ainda que o resultado da decisão desagrade o usuário.[412]

As regras procedimentais emergem, nesse contexto, com uma resposta aos problemas de legitimidade decisória das plataformas digitais, onde procedimentos claros, transparentes e participativos propiciam que os usuários aceitem as decisões

---

[410] *Ibidem*, p. 154.
[411] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 224.
[412] *Ibidem*, p. 229.

de moderação, mesmo quando lhe são desfavoráveis, conforme defende Carlos Eduardo Vieira Ramos.[413]

Para explicar o entendimento de que um procedimento transparente e participativo seria capaz de conformar os usuários mesmo quando as decisões lhe são desfavoráveis, o autor recorre à ideia de constitucionalização do mundo digital e à sociologia jurídica, para reconstruir o argumento de Niklas Luhmann na obra "Legitimidade pelo Procedimento", chamado de "Teoria dos Sistemas". Nesse sentido, veja-se abaixo a explicação do autor para a reconstrução do argumento de Luhmann:

> No capítulo anterior, uma primeira aproximação para essa solução foi feita reconstruindo o argumento de Luhmann em Legitimidade pelo Procedimento. No texto, a Teoria dos Sistemas é utilizada para explicar o papel que os procedimentos têm na estrutura política moderna – ou seja, no Estado. O argumento de Luhmann, em síntese, é de que, na sua formação, os Estados passaram a tomar decisões por procedimentos que garantem uma incerteza regrada, porque, com isso, as suas decisões se tornam legítimas – ou sejam [sic], tornam-se aceitáveis mesmo que sejam desfavoráveis para as pessoas que são impactadas por elas.[414]

O autor desenvolve, com base na Teoria dos Sistemas, a ideia de que os procedimentos são característicos por gerir "uma incerteza regrada" e podem ser compreendidos como mecanismos de "produção de decisões legítimas".[415] A incerteza regrada remete à noção de que os procedimentos "lidam com o indeferido".[416] Já a legitimidade decorre da participação daqueles que serão impactados com a decisão que será tomada ao final do procedimento. Isso significa que os potenciais atingidos por essas decisões poderão participar da construção do resultado do procedimento, o que contribui para aceitação da decisão final, independente do resultado.[417] Não à toa, a implementação de procedimentos está amplamente difundida nas sociedades modernas.

---

[413] RAMOS, Carlos Eduardo Vieira. *O Direito das plataformas: procedimento, legitimidade e constitucionalização na regulação privada da liberdade de expressão na Internet*. Dissertação (Mestrado em Direito) – Faculdade de Direito, Universidade de São Paulo, 2020, pp. 269-270.
[414] *Ibidem*, pp. 349-350.
[415] *Ibidem*, p. 294.
[416] *Ibidem*, p. 315.
[417] *Ibidem*, p. 294:"[o] procedimento caminha garantindo às suas partes a possibilidade de definir o seu destino, fazendo-as participar da construção de uma decisão final que, necessariamente, representará uma decepção para uma delas, algo que, entretanto, é esperado: faz parte do jogo processual ganhar e perder."

A partir dessa noção, Carlos Eduardo Vieira Ramos defende o desdobramento da construção original da Teoria dos Sistemas "em ferramentas capazes de enfrentar problemas que sugiram com essa nova realidade".[418] Nesse sentido, o autor entende ser necessária a adoção de técnicas procedimentais estatais pelas plataformas digitais para a solução do problema da falta de legitimidade decisória. Isso porque, a existência de um procedimento transparente e participativo aumenta a aceitação das decisões de moderação pelos usuários, bem como contribui para imunizar as decisões de moderação de conteúdo de intervenções de outros atores institucionais, como o Estado.[419]

Embora o presente trabalho de pesquisa não se filie especificamente à reconstrução da Teoria dos Sistemas de Luhmann proposta pelo autor, concorda-se com a proposta de implementação de um sistema de governança, por meio de normas procedimentais, ou melhor, de um procedimento transparente e participativo para a moderação de conteúdo das plataformas digitais.

De fato, a implementação de um procedimento como o proposto contribuiria para a regulação eficaz da liberdade de expressão no ambiente digital, assim como para o reconhecimento da legitimidade decisória e da autonomia das plataformas digitais na moderação de conteúdo, inclusive, em relação às influências externas e internas.

Partindo da premissa de que não existe um modelo de regulação perfeito ou imune a problemas, entende-se que a proposta da autorregulação regulada, por meio de normas procedimentais, é aquela que mais aproxima da construção um arcabouço regulatório capaz de promover maior transparência, credibilidade e legitimidade aos processos decisórios nas redes sociais, mesmo em face da inevitável controvérsia sobre a legalidade de determinados discursos.[420]

Portanto, em vez de definir o que pode ou não ser permitido na internet e de atribuir ao Estado o papel de fiscalização, o modelo de autorregulação regulada ou corregulação que se propõe está baseado na implementação de um procedimento de moderação de conteúdo transparente, participativo e capaz de assegurar o

---

[418] *Ibidem*, p. 323.
[419] *Ibidem*, p. 33.
[420] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 221.

cumprimento de outros deveres específicos, como do devido processo e da isonomia, conforme será desenvolvido abaixo.

3.3.1. Medidas de transparência na moderação de conteúdo

Conforme já demonstrado acima (subcapítulo 2.3.2 *supra*), medidas de transparência na moderação de conteúdo são fundamentais para a moderação de conteúdo pelas plataformas digitais, com base na aplicação de seus Termos de Uso. Elas servem para garantir que as informações necessárias sejam fornecidas para a compreensão de como as regras de moderação são aplicadas nos casos concretos. O acesso a essas informações permite a adequação do comportamento dos usuários e possibilita a pesquisa acadêmica sobre o comportamento das redes sociais.

A despeito de terem sido verificado esforços das plataformas nos últimos anos, por meio da autorregulação, em garantir maior transparência aos sistemas de moderação de conteúdo, o problema da falta de transparência das plataformas privadas persiste. Esse problema dificulta a análise precisa quanto à extensão da preocupação que se deve ter em relação à regulação privada do discurso público e aos riscos de censura prévia e censura colateral.

A melhor solução seria, evidentemente, que as próprias plataformas garantissem mecanismos transparentes, consistentes e uniformes entre si, que pudessem fornecer aos usuários as informações para a adequada compreensão dos sistemas de moderação de conteúdo, possibilitando, ainda, que esses sistemas fossem analisados por estudiosos do tema. No entanto, não sendo essa a realidade que se apresenta, é preciso pensar na intervenção regulatória estatal, por meio da autorregulação regulada, para endereçar o problema.

A correta identificação das regras de transparência é relevante para examinar os propósitos da legislação que pretende implementá-las. No subcapítulo 2.3.1.1 *supra*, demonstrou-se que a transparência cumpre funções diversas e complementares. De modo geral, a transparência serve para revelar ao público *quais* são as regras de moderação de conteúdo criadas pelas redes sociais e *quão* bem essas regras são por elas cumpridas. Além disso, o acesso à informação também permite que pesquisadores e o público identifiquem os efeitos que a operação das plataformas tem sobre as variáveis sociais, incluindo a prevenção da disseminação

de discursos de ódio, a preservação e promoção da liberdade de expressão e a influência nos processos políticos.[421]

As obrigações de transparência a serem cumpridas pelas plataformas digitais devem, portanto, ser endereçadas em três camadas: *(i)* os usuários; *(ii)* ao público em geral; e *(iii)* ao governo e/ou pesquisadores habilitados. Essa abordagem, segundo a qual algumas informações são fornecidas ao grande público, enquanto outras ficam restritas a pessoas sujeitas a deveres de confidencialidade – que, caso descumpridos, podem levar à responsabilização –, contribui para a promoção da transparência, sem que, no entanto, sejam violados a privacidade dos usuários e os sigilos empresariais das plataformas digitais.[422]

O primeiro nível de informação deve ser disponibilizado diretamente aos usuários, para que eles possam compreender os sistemas de moderação de conteúdo das plataformas que utilizam e manejar os mecanismos de reclamação, apelação ou reparação que as redes sociais oferecem. O segundo nível, por sua vez, refere-se à informação que deve ser divulgada em relatórios de acesso ao público em geral, que, aliás, são atualmente divulgados pelas redes sociais. Por fim, o terceiro nível consiste na informação sobre o funcionamento das plataformas, que deve ser divulgada apenas aos reguladores/legisladores e pesquisadores independentes previamente aprovados ou habilitados, para possibilitar a auditoria dos sistemas de moderação de conteúdo.[423]

Em atendimento aos deveres de transparência a serem observados, Mark Maccarthy elenca, ao menos, quatro recomendações gerais para os legisladores e para a própria indústria. A primeira recomendação se refere a melhorias na divulgação pública do funcionamento dos programas de moderação de conteúdo das plataformas, incluindo *(a)* regras de conteúdo em termos de uso ou normas de comunidade; *(b)* técnica de aplicação, com a remoção, despromoção e o atraso no compartilhamento de conteúdo; *(c)* procedimento para o público reclamar de possíveis violações de regras de moderação; *(d)* como as plataformas explicam suas

---

[421] MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*…, p. 8.
[422] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 247.
[423] MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*…, p. 10.

decisões às partes afetadas pelas decisões de moderação; e *(e)* procedimentos para recursos contra decisões de aplicação das regras de moderação.[424]

A segunda consiste na divulgação de relatórios periódicos às agências governamentais e ao público em geral com estatísticas agregadas que reflitam com exatidão o funcionamento dos sistemas de moderação de conteúdo. Já a terceira tem relação com a necessidade de fornecimento de termos de referência técnicos para os algoritmos utilizados na moderação de conteúdo.[425] E a quarta recomendação está relacionado ao acesso qualificado aos dados das plataformas por legisladores e pesquisadores independentes autorizados, a fim de permitir auditorias regulares de suas operações, para verificar se estão sendo cumpridas as regras preestabelecidas, devendo incluir dados relevantes do funcionamento dos sistemas de moderação.[426]

Esquematizando a divulgação dessas informações a serem reveladas dentro dos três níveis acima delineados – usuários, público em geral e legisladores ou pesquisadores independentes autorizados –, o autor recomenda que as obrigações de transparência perante os *usuários* devem incluir:[427]

(i) *Regras da plataforma*: publicação dos termos de uso e regras de comunidade, assim como das diretrizes internas que norteiam os moderadores humanos na tomada de decisões nos casos concretos, para garantir a adequada compreensão do racional das regras estabelecidas e de sua aplicação prática, além da comunicação clara e tempestiva de alterações nos termos de uso ou nas diretrizes internas;

(ii) *Ferramentas de aplicação das regras*: divulgação de informações adequadas sobre as ferramentas disponíveis em caso de violação das regras de moderação de conteúdo, para prevenir tratamentos arbitrários e ajudar a expor eventual tratamento diferenciado aplicado a determinados grupos de usuários;

(iii) *Procedimentos de revisão*: explicação do funcionamento do procedimento de revisão após o recebimento da reclamação (*flagging*), em especial em quais casos a revisão é feita de forma automatizada ou

[424] *Ibidem*, p. 15.
[425] *Idem*.
[426] *Ibidem*, p. 16.
[427] *Ibidem*, pp. 18-19.

por revisores humanos, além de fornecer informações que possibilitem o acompanhamento das reclamações e do resultado das decisões;

(iv) *Notificações*: indicação expressa do dispositivo específico das regras de moderação que teria sido violado, expondo as razões da decisão de moderação ao notificar o usuário que teve seu conteúdo removido ou cuja conta foi suspensa ou removida; e

(v) *Recursos*: fornecimento aos usuários que recorram das decisões de moderação a oportunidade de explicar por que seu conteúdo não viola as regras de comunidade indicadas na notificação, devendo a plataforma considerar, na análise desses recursos, as razões invocadas pelos usuários;

Em relação às informações a serem divulgadas pelas plataformas digitais ao *público em geral*, Maccarthy recomenda que os relatórios públicos de transparência incluam as seguintes informações:[428]

(i) *Precisão das ferramentas automatizadas e da revisão humana*: divulgação de informações sobre o índice de precisão dos mecanismos automatizados e das decisões de revisão humana para cada tipo de violação às regras da plataforma, para fornecer ao público o quadro completo da efetividade das medidas de aplicação dessas regras;

(ii) *Reporte com a extensão das violações cometidas*: fornecimento de informações sobre o número de publicações que violam os termos de uso em comparação com o número total de postagens visualizadas por usuários, para identificação do real alcance das publicações violadoras das regras da plataforma, a qualidade do conteúdo postado, bem como a forma como o conteúdo irregular é disseminado, em comparação com outros conteúdos;

(iii) *Reporte de ações de moderação adotadas:* fornecimento de informações sobre *(a)* as medidas de moderação tomadas, discriminadas por tipo de medida adotada, para que seja possível compreender a sua propensão para utilizar uma ação de aplicação severa, como a exclusão de uma

[428] *Ibidem*, pp. 19-22.

conta, em contraste com uma ação mais branda, como a despromoção de determinado conteúdo; *(b)* o número de ações adotadas em termos percentual em relação a todas mensagens ou contas que envolvam violações às regras da plataforma, para que seja viável avaliar a eficácia do esforço de moderação e da importância relativa das diferentes técnicas utilizadas; e *(c)* as medidas tomadas em função do número de utilizadores ou contas envolvidas, descontando as contas falsas, para saber se a fonte do conteúdo violador decorre de uma grande percentagem de utilizadores ou contas ou se é uma pequena fração de utilizadores ou contas que cria a maior parte do problema; e

(iv) *Medidas de efetividade dos mecanismos de moderação*: divulgação da quantidade total de conteúdos que violam as suas regras em comparação com a quantidade de conteúdos violadores detectados proativamente, antes da denúncia de usuários (*flagging*), para que seja possível atestar a efetividade dos mecanismos de moderação da plataforma.

Por último, o autor recomenda o acesso a dados relevantes sobre as operações de moderação de conteúdo das plataformas, por meio de um sistema seguro, aos reguladores e pesquisadores autorizados, para que eles possam analisar e auditar os sistemas de moderação de conteúdo, bem como conduzir pesquisas de interesse público, sem violar a privacidade dos usuários ou comprometer o valor dos dados agregados da rede social. Esse processo de divulgação de dados deveria ocorrer sob a supervisão de um órgão de corregulação independente, com atribuição para definir as pesquisas prioritárias, organizar o processo de habitação de pesquisadores e resolução de disputas entre a plataforma e os pesquisadores que decorram do acesso a essas informações.[429]

Quanto aos dados a serem disponibilizados, os legisladores e pesquisadores autorizados devem ter acesso, no mínimo, às notificações recebidas de usuários, ao conteúdo dessas notificações, às respostas das plataformas, às ações tomadas ou não tomadas diante das reclamações dos usuários, à alegada violação as regras da plataforma e ao tempo de resposta à reclamação, bem como se foi ou não requisitada uma segunda revisão e o resultado de eventual revisão. Esses dados devem ser

[429] *Ibidem*, pp. 22-23.

anonimizados para preservar a privacidade dos usuários, e os usuários do sistema devem assumir a obrigação contratual de não tentar identificar os usuários envolvidos nas reclamações, sob pena de perda do acesso ao sistema de dados e responsabilização.[430]

Todas essas recomendações, feitas por Mark Maccarthy, estão baseadas em sugestões de grupos que analisam as atuais práticas de transparência das plataformas digitais, incluindo a Comissão Europeia, o *Institute for Strategic Dialogue* e o *Data Transparency Group*, bem como se beneficiaram das medidas de devido processo propostas nos Princípios de Santa Clara.[431]

Para que essas medidas de transparência não se mostrem insuficientes, é fundamental que a regulação preveja a necessidade de padronização na divulgação das informações e os dados entre as próprias plataformas, evitando que a variedade e o detalhamento das informações e das classificações dos dados divulgados criem obstáculos à realização de análises comparativas. Do contrário, apesar dos esforços para garantir maior transparência, não será possível comparar o desempenho entre plataformas ou entender a variação na aplicação das políticas entre elas.[432]

3.3.2. Devido processo e isonomia

Embora as medidas de transparência sejam reconhecidamente as mais relevantes para a regulação das redes sociais, justamente por ser uma condição necessária para as demais garantias procedimentais, deve-se defender a observância pelas plataformas digitais dos direitos ao devido processo e à isonomia.[433]

O devido processo e a isonomia são direitos processuais constitucionais (CF, artigo 5°, *caput* e inc. LIV), que estão também consagrados padrões de liberdades individuais e princípios internacionais de direitos humanos, incluindo documentos de *softlaw* como os Princípios de Manila e os Princípios de Santa Clara, conforme visto anteriormente (subcapítulo 2.3.2 *supra*).

Os Princípios de Manila recomendam, em seu quinto princípio, que as práticas de moderação devem respeitar o devido processo e, no terceiro princípio,

---

[430] *Ibidem*, p. 24.
[431] *Ibidem*, p. 16.
[432] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 252.
[433] *Idem*.

que as solicitações para restrição de conteúdo devem ser claras e seguir o devido processo, incluindo o direito de o usuário ser ouvido antes de qualquer restrição de conteúdo.[434] Dentre os princípios fundamentais, uma das categorias dos Princípios da Santa Clara, destaca-se ainda o dever das plataformas de integrar considerações sobre direitos humanos e devido processo em todas as etapas da moderação de conteúdo e de divulgar informações sobre essa integração.[435]

Para o cumprimento do devido processo, é preciso que as plataformas digitais informem aos usuários que foram afetados pelas decisões de moderação de conteúdo qual o dispositivo dos termos de uso ou das regras de comunidade que teria sido violado e, portanto, justificado a medida de moderação.[436]

Além disso, as plataformas devem garantir aos usuários a apresentação de reclamações ou recursos, em sistema próprio, contra decisões de remoção ou bloqueio, suspensão ou cessação da prestação do serviço, suspensão ou encerramento de conta, desmonetização, redução do alcance ou associações de mensagem de esclarecimentos ou checagem de fatos.[437] Também é importante que as plataformas disponibilizem aos usuários um sistema específico para que os usuários possam apresentar denúncias (*flags*) devidamente fundamentadas de conteúdos ou contas de outros usuários por violações de regras das plataformas ou leis locais.[438]

É recomendável que as decisões sobre as reclamações ou recursos apresentados pelos usuários sejam tomadas dentro de prazo razoável, sem demora injustificada, mas considerando, evidentemente, o volume de apelações apresentadas às plataformas digitais dentro daquele período em que a reclamação fora apresentada pelo usuário. Essas decisões proferidas em sede recursal não deverão ser tomadas de forma exclusivamente automatizada.[439]

Em relação aos deveres de isonomia, ainda que existam critérios diversos, por exemplo, para pessoas públicas ou informações de interesse público, a adoção

---

[434] ELETRIC FRONTIER FOUNDATION *et al*. *Princípios de Manila sobre Responsabilidade de Provedores*. Disponível em: <https://manilaprinciples.org/index.html>. Acesso em 17 fev. 2025

[435] ACESS NOW *et al*. Os *Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo*. Disponível em: <https://santaclaraprinciples.org/>. Acesso em 17 fev. 2024.

[436] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*..., p. 272.

[437] *Idem*.

[438] *Ibidem*, p. 273.

[439] *Idem*.

de critério decisório não isonômico para a moderação de conteúdo deve ter caráter excepcional. Essa questão, inclusive, já foi objeto de controvérsias entre o Comitê Supervisor (*Oversight Board*) do *Facebook* e a Meta (empresa detentora do *Facebook*). Ao final do primeiro ano de seu funcionamento, o Comitê Supervisor publicou relatório de transparência identificando que determinados usuários – incluindo o atual Presidente americano, Donald Trump – detinham maior liberdade na plataforma do que outros.[440]

Delineadas acimas as garantias procedimentais a serem implementadas no contexto da moderação de conteúdo das plataformas digitais, por meio do modelo da autorregulação regulada, baseado na implementação de normas processuais, passa-se, então, a analisar as principais propostas de organizações nacionais e internacionais sobre a moderação de conteúdo por meio do estabelecimento de deveres procedimentais às plataformas.

3.3.3. Exame das principais propostas de organizações nacionais e internacionais sobre a moderação por normas de procedimento

A moderação de conteúdo por normas procedimentais tem sido proposta há mais de uma década por entidades não governamentais ao redor do mundo. Com o início do debate no ordenamento jurídico brasileiro, instituições públicas e privadas comentaram sobre a legalidade e eficiência do devido processo na moderação de conteúdo, gerando a expectativa de que as análises feitas fossem encontradas com mais facilidade na internet.

O cenário encontrado, contudo, foi outro. Os comentários da sociedade civil não são encontrados de forma organizada em qualquer das casas legislativas em que tramitaram as propostas de regulação das redes sociais. [441] Não bastasse a falta de compilação dos estudos, é notável como as regras procedimentais recebem tratamento um tanto superficial, que se limita a apresentar as exigências mínimas em torno da notificação da remoção de conteúdo, a possibilidade de recorrer da

[440] *Idem*.

[441] Um cenário de difícil compreensão sobre a evolução do PL 2.630 foi a profusão de versões e suas respectivas divergências disponibilizados pelo governo federal e a do próprio relator. As diferenças foram identificados pelo ITS RIO. *Tabela comparativa das versões do PL 2630*, publicado pelo Vozes da regulação. Disponível em: <https://www.vozesdaregulacao.org.br/analise-pl2630>. Acesso em 13 fev. 2025.

decisão das redes sociais e, ao final, a exigibilidade do relatório para satisfazer o dever de transparência sobre a moderação do conteúdo público.

Por causa disso, a análise abaixo aborda, primeiro, as entidades que alcançaram maior projeção no debate nacional sobre a matéria, com suas versões mais atualizadas, e, segundo, as propostas diretrizes e/ou comentários de instituições internacionais que possam efetivamente colaborar com as propostas normativas no país.[442]

É importante destacar que, embora diversas entidades relevantes tenham contribuído para o debate, este trabalho se concentra naquelas que, dentro de seu escopo, apresentaram análises ou posições específicas sobre normas procedimentais na moderação de conteúdo. Essa escolha metodológica não desconsidera outras contribuições igualmente significativas, mas reflete a necessidade (prática) de maior foco na regulação das redes sociais no viés procedimental. Após sua análise, foram compilados os institutos e as estruturas mais importantes para o controle da moderação de conteúdo.

---

[442] PL 2.630/2020, redação atual: "Art. 16.: Os provedores deverão criar mecanismos que permitam a qualquer usuário notificá-los da presença, em seus serviços, de conteúdos potencialmente ilegais, de forma justificada. § 1º O mecanismo e os requisitos mínimos para a notificação de conteúdos serão definidos em regulamento. § 2º O registro da notificação de que trata este artigo configura-se como ato necessário e suficiente como prova do conhecimento pelos provedores sobre o conteúdo apontado como infringente, para fins do disposto no art. 13 desta lei. Art. 17. O procedimento de moderação de conteúdo e de conta deve observar o normativo vigente e ser aplicado com equidade, consistência e respeito ao direito de acesso à informação, à liberdade de expressão e à livre concorrência. Parágrafo único. Os termos de uso, quanto à moderação de conteúdo e de contas, devem sempre estar orientados pelos princípios da necessidade, proporcionalidade e não discriminação, inclusive quanto ao acesso dos usuários aos serviços dos provedores. Art. 18. Após aplicar as regras contidas nos termos de uso que impliquem moderação de conteúdos, incluindo aquelas envolvendo alteração de pagamento monetário ou publicidade de plataforma, os provedores de redes sociais e de mensageria instantânea devem, ao menos: I - notificar o usuário que publicou o conteúdo sobre: a) a natureza da medida aplicada e o seu âmbito territorial; b) a fundamentação, que deve necessariamente apontar as cláusulas de seus termos de uso para aplicação e o conteúdo ou a conta que deu causa à decisão; c) 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, nos termos do art. 20, § 1º, da Lei nº 13.709, de 14 de agosto de 2018, quando cumpridos os requisitos para tanto; II - responder de modo fundamentado e objetivo aos pedidos de revisão de decisões e providenciar a sua reversão imediata quando constatado equívoco. § 1º O código de conduta deverá dispor sobre os prazos razoáveis para cumprimento dos incisos I e II deste artigo. § 2º Em caso de provimento do pedido de revisão, as medidas aplicadas devem ser imediatamente revogadas, devendo ser dada publicidade ao equívoco constatado. § 3º Disponibilizar, por um prazo mínimo de seis meses, canal próprio destacado e de fácil acesso para formulação de denúncias sobre conteúdos e contas em operação e envio de pedido de revisão de decisões e consulta do histórico de interações entre o provedor e o usuário. Art. 19. Os provedores de que trata esta Lei devem: I - criar mecanismos para informar publicamente a ação, pelo provedor, de moderação de conteúdo, independente da causa que deu origem à moderação; e II - manter pública a identificação de ação judicial que deu origem à moderação em conteúdos, ressalvados processos em sigilo".

3.3.3.1. Coalizão Direitos nas Redes (“CDR”)

A Coalizão Direitos nas Redes (“CDR”)[443] defende que, quando as plataformas digitais receberem solicitações de moderação de perfis de usuários e seus conteúdos, o usuário que poderá vir a ser afetado deverá ser notificado a respeito da solicitação, para que seja ouvido antes da decisão de moderação.[444] Em vista da gravidade de alguns conteúdos, o CDR entende que o dever de notificar o usuário pode ser excepcionado em determinadas hipóteses.[445]

Essas hipóteses não devem ser engessadas em um rol taxativo. Ao invés disso, a entidade sugere que as exceções ao dever de notificar sejam elaboradas por uma entidade fiscalizadora independente, que, em sua atividade regulatória, efetue a atualização do rol a cada dois anos, considerando a constante evolução no ambiente digital.[446_447] Essa atualização deve ter por base as demandas identificadas pelo corpo técnico da nova entidade reguladora/fiscalizadora.[448]

---

[443] O CDR reúne mais de 40 organizações da sociedade civil, da academia e ativistas para defender os direitos humanos na Internet.

[444] COALIZÃO DIREITOS NAS REDES. *PL 2630: propostas da CDR para uma lei efetiva e democrática*, pp. 14-15. Disponível em: <https://direitosnarede.org.br/2020/09/01/pl-2630-propostas-da-cdr-para-uma-lei-efetiva-e-democratica/. Acesso em 10 fev. 2025.

[445] *Ibidem*, p. 15.

[446] BRASIL. *Projeto de Lei 2.630*. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=Tramitacao-PL%202630/2020. Acesso em 07 mar. 2025: “Art. 25. O Congresso Nacional instituirá, em até 60 (sessenta) dias contados da publicação desta Lei, em ato próprio, conselho que terá como atribuição a realização de estudos, pareceres e recomendações sobre liberdade, responsabilidade e transparência na internet. Parágrafo único. O Conselho de Transparência e Responsabilidade na Internet é o órgão responsável pelo acompanhamento das medidas de que trata esta Lei, e a ele compete: (...) II – elaborar código de conduta para redes sociais e serviços de mensageria privada, a ser avaliado e aprovado pelo Congresso Nacional, aplicável para a garantia dos princípios e objetivos estabelecidos nos arts. 3º e 4º desta Lei, dispondo sobre fenômenos relevantes no uso de plataformas por terceiros, incluindo, no mínimo, desinformação, discurso de incitação à violência, ataques à honra e intimidação vexatória; (...).”

[447] COALIZÃO DIREITOS NAS REDES. *PL 2630: propostas da CDR para uma lei efetiva e democrática...*, p. 26: “II – elaborar código de conduta para redes sociais, ferramentas de busca e serviços de mensageria privada, revisado a cada 2 (dois) anos, visando a adequação de suas políticas de uso aos princípios e objetivos estabelecidos na Lei, assim como a garantia de sua consecução; (...)”.

[448] Em defesa da proposta, a CDR expõe que “[c]onsiderando o caráter técnico do Conselho de Transparência e Responsabilidade na Internet, o código de conduta que será elaborado pelo mesmo não deve ser submetido à aprovação do Congresso Nacional, o que conferiria status de norma infralegal a um documento que, devido à dinâmica da evolução tecnológica, deve ser revisado a cada dois anos. A medida também possibilitaria eventual revisão e ingerência política nas decisões de um conselho técnico multissetorial. Por fim, não cabe à lei, de antemão, determinar de que deve tratar o código de conduta.” (*Idem*).

Ao final dos comentários sobre as regras procedimentais, a CDR afirma que os usuários devem ter direito à instância revisional, sem desenvolver em detalhes, no entanto, a estrutura do órgão ou requisitos para a revisão da decisão de moderação de conteúdo.[449]

Os demais comentários se relacionam à reparação dos conteúdos lícitos moderados indevidamente ou dos conteúdos ilícitos não moderados. Segundo a CDR, não seria uma reparação pecuniária, mas sim uma reparação pela remoção indevida do conteúdo lícito ou pela não remoção de conteúdo ilícito, dentro do alcance técnico das plataformas digitais, que seriam estabelecidos em um código de conduta a ser elaborado por entidade reguladora/fiscalizadora independente.[450]

Nesse aspecto, a CDR conclui que o direito de resposta dos usuários afetados exigiria decisão judicial, indicando que a plataforma deve viabilizar um alcance proporcional ao agravo. E que, em caso de danos oriundos de conteúdo publicitário, o causador do dano ficaria responsável pelo pagamento do impulsionamento do direito de resposta.[451]

A CDR também advoga pela extinção do privilégio procedimental às autoridades e figuras públicas quando forem vítimas de *deep fake*. Em sua perspectiva, todos os temas afetos à moderação de conteúdo, aí inclusa as *deep fakes*, deveriam ser tratados tecnicamente no âmbito da entidade reguladora, e não em uma previsão legislativa rígida e imutável.[452]

Em comentários mais recentes, a CDR afirma que, embora a doutrina do *Notice and Takedown* implemente a moderação de conteúdo em alguns países,

---

[449] *Idem*.

[450] Proposta: "§ 4º Havendo dano decorrente da caracterização equivocada de conteúdos como violadores dos padrões de uso de aplicações ou do disposto na presente Lei, caberá ao provedor de redes sociais promover medidas não-pecuniárias de reparação, no âmbito e nos limites técnicos do serviço, de acordo com termos definidos no código de conduta previsto no art. 25." (*Ibidem*, pp. 15-16).

[451] Proposta: "Art. XX - Diante de decisão judicial, cabe às redes sociais, no âmbito e nos limites técnicos dos seus serviços, garantir alcance, proporcional ao agravo, do direito de resposta, recaindo sobre o responsável, quando se tratar de impulsionamento e publicidade, as expensas pela divulgação do conteúdo." (*Idem*).

[452] *Ibidem*, p. 16.

existe um amplo movimento que confirma a necessidade de aplicação do devido processo à moderação de conteúdo.[453_454]

3.3.3.2. Instituto de Tecnologia & Sociedade do Rio – ITS Rio

Em março de 2022, o Instituto de Tecnologia & Sociedade do Rio – ITS Rio disponibilizou ao público relatório com pontos de atenção sobre, na época, a principal proposta de regulação das redes sociais no contexto brasileiro.[455_456] A entidade afirmou que a principal proposta legislativa de moderação de conteúdo, o PL da Fake News, desenvolveu "um mini Código de Processo Civil", para implementar um procedimento, que compreenderia:

> *(i)* *Notificações*: envio de notificação de decisões de moderação de conteúdo ao usuário sobre conteúdos ou contas moderadas;[457]

---

[453] COALIZÃO DE DIREITOS NAS REDES. *Relatório de Referências Internacionais em regulação de plataformas digitais: bons exemplos e lições para o caso brasileiro*, publicado em 23 abr. 2024. Disponível em: <https://direitosnarede.org.br/2024/04/23/coalizao-direitos-na-rede-lanca-o-relatorio-referencias_internacionais-em-regulacao-de-plataformas-digitais-bons-exemplos-e-licoes-para-o-caso brasileiro/>. Acesso em 10 fev. 2025.

[454] Nesse sentido: "Apesar deste recuo grave, o projeto segue trazendo regras importantes para limitar o poder das plataformas digitais e empoderar a sociedade e, portanto, deve prosperar. É o caso das obrigações de transparência, atenção aos termos de uso e outras políticas das plataformas, bem como das regras do chamado devido processo (como exigências de notificação do usuário quando da moderação de conteúdo e de mecanismos de recurso). O projeto é fundamental, portanto, para que possamos conhecer mais como funcionam espaços que se tornaram extremamente relevantes para o debate público e para envolver a sociedade na busca para que eles sejam sadios, por isso há proposições de mecanismos para denúncias de conteúdos criminosos e acesso a informações." (COALIZÃO DIREITOS NAS REDES. PL 2630: Regulação pública democrática das plataformas é fundamental, com instituições autônomas e participativas, publicado em: 28 abr. 2023. Disponível em: <https://direitosnarede.org.br/2023/04/28/pl-2630-regulacao-publica-democratica-das-plataformas-e-fundamental-com-instituicoes-autonomas-e-participativas/>. Acesso em 14 fev. 2025).

[455] ITS RIO. *9 pontos de atenção sobre a PL das Fake News (PL 2630/2020)*, publicado em 31 mar. 2022. Disponível em: <https://itsrio.org/wp-content/uploads/2022/04/9-pontos-de-aten%C3%A7%C3%A3o-sobre-o-PL-das-Fake-News-PL-2630_20.pdf>. Acesso em 14 fev. 2025.

[456] Em 28 de março de 2022, o ITS Rio havia publicado uma versão antiga do documento, indicando dez pontos de atenção. Naquela versão houve uma supressão do ponto que tratava da regulação da internet como veículo de comunicação social, presente no segundo ponto de atenção. Não houve uma explicação explícita sobre o porquê da mudança. (ITS RIO. *10 pontos de atenção sobre a PL das Fake News (PL 2630/2020)*, publicado em 28 mar. 2022. Disponível em: <https://itsrio.org/wp-content/uploads/2022/03/10-pontos-de-atencao-sobre-o-PL-das-Fake-News-PL-2630_20.pdf>. Acesso em 13 fev. 2025).

[457] ITS RIO. *9 pontos de atenção sobre a PL das Fake News (PL 2630/2020)...*, p. 4: "Empresas terão regras de processo para moderar conteúdo, com direito de resposta e informação sobre o perfil de quem modera (art. 15). A construção de regras procedimentais para a moderação de conteúdo é uma tendência global. O PL2630 traz regras sobre notificação ao usuário sobre conteúdos e contas moderadas, além de dispor sobre pedidos de revisão dessas decisões."

(ii) *Reparação*: direito de resposta do usuário para que possa haver desagravo público acerca das informações e/ou opinião ilegalmente emitidas;

(iii) *Recursos*: direito do usuário à revisão das decisões de moderação de conteúdo sobre conteúdos e contas moderadas;[458]

(iv) *Relatórios de Transparência*: envio de relatórios semestrais com as características gerais das equipes envolvidas na aplicação de termos e políticas de uso em relação a conteúdos gerados por terceiros, o número de pessoas envolvidas na atividade, o modelo de contratação, além de estatísticas sobre seu idioma de trabalho, qualificação, indicativos de diversidade atributos demográficos e nacionalidade;[459] e

(v) *Órgão fiscalizador*: atribuição de novas competências ao Comitê Gestor da Internet – CGI.br, podendo formular diretrizes para elaboração e validação de códigos de conduta, avaliar relatórios de transparência e requisitar informações sobre metodologias de moderação de conteúdo.[460]

O relatório não analisou detalhadamente as obrigações procedimentais da proposta legislação, deixando de tecer comentários sobre o conteúdo mínimo da notificação a ser enviada ao usuário, a estrutura ou funcionamento do órgão fiscalizador ou mesmo sobre o mecanismo para o exercício do direito de resposta do usuário afetado pela decisão de moderação de conteúdo.

---

[458] *Idem*.

[459] *Idem*: "O PL inova ao exigir que as empresas informem semestralmente as "características gerais das equipes envolvidas na aplicação de termos e políticas de uso em relação a conteúdos gerados por terceiros, incluindo número de pessoas envolvidas na atividade, modelo de contratação, bem como estatísticas sobre seu idioma de trabalho, qualificação, indicativos de diversidade atributos demográficos e nacionalidade."

[460] *Ibidem*, p. 5: "O Comitê Gestor da Internet (CGI.br) ganha diversas novas competências com o PL2630, como a formulação de diretrizes para elaboração e a validação de Códigos de Conduta para os provedores de redes sociais, ferramentas de busca e aplicativos de mensagem. Ao CGI também caberá avaliar os relatórios de transparência semestrais produzidos pelos provedores. Vale destacar que o CGI passaria também a poder requerer diretamente aos provedores informações sobre metodologias de moderação de conteúdo, procurando esclarecer como e porquê contas e conteúdos foram excluídos, desindexados ou sinalizados como falso ou enganoso, por exemplo."

3.3.3.3. Internet Lab

Em agosto de 2021, o Internet Lab lançou a 5ª edição do "Diagnósticos & Recomendações", com o objetivo de colaborar com o debate acerca da moderação de conteúdo no Brasil. O relatório deu ênfase na perspectiva procedimental da moderação de conteúdo, criticando os riscos criados por propostas que tem como mote a legalidade das manifestações.[461] Em vista disso, o Internet Lab propôs as seguintes seções:

(i) *Notificações*: envio de notificação aos usuários afetados pela decisão de moderação de conteúdo;

(ii) *Recursos*: instância recursal para revisão da decisão de moderação de conteúdo tomada pela plataforma digital; e

(iii) *Relatórios de transparência*: disponibilização de informações acessíveis sobre o procedimento de moderação de conteúdo aos usuários, bem como de relatórios periódicos de transparência a respeito da moderação de conteúdo.

De acordo com o Internet Lab, os usuários devem receber notificação com informações detalhadas sobre o conteúdo objeto da decisão de moderação, contendo: "*(a)* apresentação da URL ou de um excerto do conteúdo removido; *(b)* a indicação específica da política ou termo de serviço violado; *(c)* a forma como a violação foi identificada — denúncia de autoridades, sistemas automatizados, sinalização de outros usuários ou de *trusted flaggers* [sinalizadores confiáveis]; e *(d)* a explicação clara do procedimento para recorrer da decisão."[462]

---

[461] MONTEIRO, Artur Pericles Lima; CRUZ et al. Francisco Brito; SILVEIRA, Juliana Fonteles da; e VALENTA, Mariana G. *Armadilhas e caminhos na regulação da moderação de conteúdos: Diagnósticos & Recomendações #5*. São Paulo: InternetLab, 2021, p. 21: "Mesmo quanto a conteúdo de fato ilícito, como violação da imagem e da privacidade, por exemplo, persistem tensões nas exceções à regra geral de vedação da moderação de conteúdo proposta no texto do Ministério do Turismo. O que caracteriza esses ilícitos está longe de ser incontroverso. E isso tem consequências que vão muito além. (...) O grande problema aqui se revela quando pensamos em como as plataformas reagirão diante dessas incertezas. Elas abrem espaço para responsabilização tanto caso provedores removam conteúdo que não deviam — por exemplo, porque avaliaram que uma publicação se enquadrava numa exceção à regra geral —, quanto não removam conteúdo que deviam — porque consideraram que não se aplicava uma exceção. Erros nesse juízo sobre regra geral e exceções sujeitaria plataformas e outros provedores a duras sanções, que incluiriam até mesmo o fim das operações no Brasil, segundo a minuta divulgada pelo Executivo em maio."

[462] *Ibidem*, p. 28.

As informações indicadas pela Internet Lab parecem garantir mais segurança, impedindo possíveis alegações de nulidade em torno da notificação (impossibilidade de identificar o conteúdo) ou a fonte que solicitou o procedimento de moderação de conteúdo. Assim, o usuário poderá direcionar adequadamente suas pretensões e fundamentar seus pedidos, inclusive na esfera judicial.

Em seguida, o Internet Lab ressaltou a necessidade de haver uma instância recursal para recorrer das decisões de moderação das redes sociais. O mecanismo recursal deveria incluir: "*(a)* revisão humana por alguém não envolvido na decisão original; *(b)* a possibilidade de o usuário fornecer informações adicionais para a apreciação do recurso; e *(c)* uma declaração por escrito do resultado e das razões adotadas na apreciação do recurso."[463]

O Internet Lab propôs ainda a necessidade de as plataformas divulgarem informações gerais e compreensíveis sobre políticas e sistemas de moderação. Com o objetivo de prestar contas sobre os sistemas de moderação de conteúdo, sugeriu que as redes sociais publiquem suas políticas de conteúdo de forma acessível e clara, em idioma nacional, e notifiquem os usuários em caso de atualizações.

Desse modo, as informações a serem divulgadas deveriam conter: "*(a)* o tipo de conteúdo e atividades proibidos, *(b)* as providências adotadas para cada violação; *(c)* os critérios utilizados pelos mecanismos de moderação e curadoria, considerando contextos culturais e idiomáticos; *(d)* o impacto da curadoria de conteúdo em sua visibilidade; *(e)* os critérios para a utilização de moderação humana e automatizada; e *(f)* a quantidade de moderadores alocados para a realização da atividade em nível nacional."[464]

Por fim, o Internet Lab destacou como fundamental o fornecimento de dados periódicos sobre a aplicação de políticas de moderação. Em levantamento da entidade, foi constatado que ainda faltam informações que permitam compreender o cenário em cada país ou região, já que os dados sobre moderação de conteúdo geralmente são apresentados de forma agregada em nível global.

Nesse contexto, as plataformas deveriam publicar relatórios que indiquem o número total de publicações e contas sinalizadas (*flagged*) e removidas, especificando: *(a)* o formato (vídeo, áudio, imagem, texto ou *livestream*); *(b)* a fonte

[463] *Ibidem*, p. 29.
[464] *Idem*.

da sinalização (determinação governamental, algoritmos, sinalização de outros usuários ou de *trusted flaggers*); e *(c)* os locais das remoções.[465]

3.3.3.4. Centro de Tecnologia e Sociedade da FGV (CTS da FGV)

Após examinar o regime do procedimento de moderação em debate no ordenamento jurídico brasileiro, o CTS da FGV ratificou a necessidade de ser garantido no contexto da moderação de conteúdo os seguintes direitos ao usuário:[466]

(i) *Notificação*: envio de notificação aos usuários sobre eventual restrição, remoção ou suspensão de conteúdo ou conta;
(ii) *Recursos*: revisão ou contestação da decisão de moderação; e
(iii) *Reparação*: direito de resposta em caso de retirada indevida do conteúdo lícito ou ausência de moderação tempestiva do conteúdo qualificado como ilícito.

O principal ponto de atenção apontado pelo CTS da FGV é necessidade de se disciplinar um "prazo razoável para que a existência de erros seja constatada" no contexto da moderação de conteúdo, bem como de um prazo para a conclusão do procedimento decisório das plataformas digitais.[467]

Por causa disso, o CTS propôs a necessidade de uma determinação geral para fixação de prazos, que respeitem a proporcionalidade e razoabilidade, ainda que esses prazos sejam definidos pela entidade reguladora que ficará responsável por fiscalizar a implementação da regulação das redes sociais. Nesse contexto, foi mencionada também a necessidade de ser restabelecido imediatamente o conteúdo e/ou conta alvo da moderação das plataformas caso seja reconhecido a aplicação equivocada das regras de moderação. [468]

---

[465] *Idem*.
[466] *Idem*.
[467] *Ibidem*, p. 27.
[468] CURZI, Yasmin. ZINGALES, Nicolo. GASPAR, Walter. LEITÃO, Clara. COUTO, Natália. REBELO, Leandro. OLIVEIRA, Maria Eduarda. *Nota técnica do Centro de Tecnologia e Sociedade da FGV Direito Rio sobre o substitutivo ao PL 2630/2020*. Rio de Janeiro: FGV Direito Rio, 2021, pp. 25-27.

3.3.3.5. Eletronic Frontier Foundation (EFF) e AccessNow

Em 7 de julho de 2023, a EFF e a AccessNow publicaram um estudo sobre os pontos de maior preocupação sobre as propostas normativas em torno da moderação de conteúdo, especialmente quando estabelece obrigações para provedores de aplicações em situações de risco iminente de dano ou negligência, visando fundamentar intervenções em crises.[469]

As organizações apontam para a necessidade de maior clareza e precisão na utilização de mecanismos de controle adequados para garantir intervenções necessárias e proporcionais. Nesse sentido, defenderam que o mecanismo de notificação e retirada, vinculado a um protocolo de segurança, poderia moderar pressões para expandir exceções à responsabilidade de intermediários. Sua limitação temporal e de escopo poderia mitigar preocupações, que assegura o direito de recurso contra decisões de moderação de conteúdo.

Além disso, a EFF e a AccessNow também indicaram a necessidade de as notificações especifiquem a localização do material alegadamente ilegal e justifiquem sua ilegalidade com base nas normas do ordenamento jurídico brasileiro ou nos termos de uso/políticas da comunidade. Por fim, outro ponto que, de acordo com as organizações, mereceria atenção seria a garantia ao devido processo no processo decisório de moderação de conteúdo.

3.3.3.6. Organização das Nações Unidas

Em 2023, a Unesco elaborou um guia com diretrizes para a regulação das plataformas digitais para garantir a liberdade de expressão e o acesso à informação, seguindo uma abordagem multissetorial em torno do problema.[470]

Uma das principais preocupações da Unesco foi estabelecer uma explicação detalhada em torno da garantia do devido processo na moderação de conteúdo:

---

[469] Eletronic Frontier Foundation – EFF. P*adrões de Direitos Humanos como Linhas de Base para a Regulação e Prestação de Contas das Plataformas: uma contribuição para o debate brasileiro*. Disponível em: <https://www.eff.org/files/2023/07/07/padroes_de_direitos_humanos_como_linhas_de_base_para_a_regulacao_e_prestacao_de_contas_das_plataformas_pt-br.pdf>. Acesso em 18 fev. 2025.

[470] UNESCO. *Safeguarding freedom of expression and access to information: guidelines for a multistakeholder approach in the context of regulating digital platforms*. Publicado em 27 abr. 2023. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en>. Acesso em 13 fev. 2025.

> 107. Além da plataforma digital disponibilizar informações sobre suas políticas de forma acessível e em um formato compreensível em todos os idiomas relevantes, deve também demonstrar como usuários e não usuários, ou terceiros agindo em nome de usuários e não usuários, podem relatar potenciais abusos das políticas. Há também a necessidade de mecanismos eficazes de reclamações acessíveis para crianças. As plataformas digitais também devem ter meios para compreender o contexto local e as condições locais ao responder às reclamações dos usuários e garantir que seus sistemas sejam projetados de forma culturalmente sensível. 108. O sistema de relatórios de usuários deve dar alta prioridade às preocupações relacionadas ao conteúdo que ameace os usuários, garantindo uma resposta rápida e, se necessário, fornecendo um canal específico de escalonamento ou meio de registro da denúncia. Isso é especialmente importante no que diz respeito à violência e assédio com base em gênero. 109. Deve haver um mecanismo eficaz de reparação de usuários na plataforma e externamente, permitindo aos usuários (e não usuários, se afetados por conteúdo específico) oportunidades significativas para levantar suas preocupações e obter reparação quando apropriado. Isso deve incluir um canal claro, facilmente acessível e compreensível para reclamações, e os usuários devem ser notificados sobre o resultado de seu recurso. 110. O mecanismo de apelação deve seguir os sete princípios delineados nos Princípios Orientadores da ONU sobre Empresas e Direitos Humanos para mecanismos eficazes de reclamações: legitimidade, acessibilidade, previsibilidade, equitabilidade, transparência, compatibilidade com os direitos e aprendizado contínuo. 111. As plataformas digitais devem notificar os usuários e explicar os processos de apelação quando seu conteúdo for removido ou explicitamente rotulado, restrito em termos de comentários ou compartilhamento novamente, ou associado à publicidade, com limites especiais em termos de amplificação ou recomendação (distintos da amplificação e recomendação "orgânica/algorítmica"). Isso permitiria aos usuários entender os motivos pelos quais ação foi tomada em seu conteúdo, o método utilizado (algorítmico ou após revisão humana) e sob quais regras da plataforma a ação foi tomada. Além disso, elas devem ter processos em vigor que permitam aos usuários recorrer dessa decisão e obter o devido reparo. [471]

[471] *Ibidem*, pp. 28-29: “107. In addition to the digital platform making information about its policies accessible in a digestible format and in all relevant languages, it should demonstrate how users and non users, or third parties acting in the interests of users and non users can report potential abuses of the policies. There is also a need for effective user complaints mechanisms that are accessible for children. Digital platforms should also have the means to understand local context and local conditions when responding to user complaints and ensure that their systems are designed in a culturally sensitive way. 108. The user reporting system should give high priority to concerns regarding content that threatens users, ensuring a rapid response, and, if necessary, by providing a specific escalation channel or means of filing the report. This is particularly important when it to comes to gender-based violence and harassment. 109. There should be an effective on-platform and external user redress mechanism to allow users (and non-users if impacted by specific content) meaningful opportunities to raise their concerns and secure redress where appropriate. This should include a clear, easily accessible, and understandable reporting channel for complaints, and users

Em síntese, as diretrizes estabelecidas pela Unesco indicam que as plataformas digitais devem cumprir com os seguintes requisitos, para atender ao cumprimento dos aludidos deveres procedimentais:

(i) *Informações acessíveis*: as redes sociais devem explicar de forma didática e contínua sobre como o procedimento de moderação de conteúdo tem sido conduzido e as diretrizes utilizadas para a tomada de decisões nesse contexto;

(ii) *Mecanismos de denúncia (flagging)*: oferecer meios para que usuários denunciem conteúdos abusivos ou ilegais, garantindo respostas dentro de prazo razoável;

(iii) *Resposta a reclamações*: disponibilizar canais para responder às reclamações dos usuários de forma eficiente;

(iv) *Notificação e explicação*: notificar os usuários e explicar os processos de apelação quando seu conteúdo for removido ou rotulado, detalhando os fundamentos da decisão de moderação;

(v) *Recursos*: estabelecer processos que permitam aos usuários recorrerem de decisões de moderação de conteúdo e obterem reparação adequada em caso de falhas; e

(vi) *Relatórios de transparência*: disponibilizar em relatórios públicos o número de recurso e natureza dos solicitantes por moderação de conteúdo.

A Relatora Especial sobre Liberdade de Expressão da ONU, Irene Khan, destaca que propostas regulatórias focadas em transparência e no cumprimento das obrigações do devido processo podem contribuir positivamente para a proteção dos

---

should be notified about the result of their appeal. 110. The appeals mechanism should follow the seven principles outlined in the UN Guiding Principles on Business and Human Rights for effective complaints mechanisms: legitimacy, accessibility, predictability, equitability, transparency, rights-compatibility, and continuous learning. 111. Digital platforms should notify users and explain processes for appeal when their content is removed or expressly labelled, restricted in terms of comments or re-sharing or advertising association, given special limits in terms of amplification or recommendation (as distinct from "organic/algorithmic" amplification and recommendation), and why. This would allow users to understand the reasons that action on their content was taken, the method used (algorithmic or after human review), and under which platform rules action was taken. Also, they should have processes in place that permit users to appeal such decision and have appropriate redress."

direitos humanos e para uma maior responsabilidade pública das plataformas.[472] Diante disso, a relatora aponta a necessidade de construir um sistema recursal que permita que as decisões sobre moderação possam ser recorridas pelas pessoas afetadas ou interessadas no conteúdo.[473]

3.3.3.7. Código de Conduta contra Desinformação de 2022 da Comissão Europeia

Em 2018, a Comissão da União Europeia elaborou o *Código de Conduta sobre Desinformação*, por meio do qual atores do setor digital concordaram em estabelecer padrões autorregulatórios para combater a desinformação *online*. Esse código se tornou um importante pilar na estratégia da União Europeia contra a desinformação, demonstrando sua eficácia especialmente durante períodos eleitorais e crises, como a pandemia de COVID-19 e a guerra na Ucrânia.

Após uma avaliação da Comissão Europeia sobre sua implementação inicial, foi publicada, em maio de 2021, uma orientação detalhada para corrigir deficiências do Código de 2018 e torná-lo mais eficaz. Esse processo levou à revisão do Código, que foi apresentado à Comissão em 16 de junho de 2022 por 34 signatários.[474] Em 13 fevereiro de 2025, o Código foi formalmente integrado ao

---

[472] KHAN, Irene. *Disinformation and freedom of opinion and expression. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. United Nations General Assembly, Human Rights Council, Forty-seventh session, A/HRC/47/25. Publicado em 13 abr. 2021. Disponível em: <https://digitallibrary.un.org/record/3925306>. Acesso em 13 fev. 2025.

[473] *Ibidem*, p. 15: "72. Companies continue to fail to provide adequate remedies for wrongful actions taken on the basis of disinformation or misinformation. Appeals mechanisms for wrongful decisions are crucial to offset the significant risks inherent in large social media companies using imperfect filters to remove content. Appeals do not appear, however, to be available for enforcement actions taken by companies such as labelling and demotions. Nor do they appear to be available to challenge decisions taken on the basis of coordinated harm or inauthentic behavior policies. Moreover, it is unclear whether appeals mechanisms are available in a range of languages". Tradução: "As empresas continuam a falhar em fornecer soluções adequadas para ações indevidas tomadas com base em desinformação ou informação incorreta. Mecanismos de apelação para decisões equivocadas são cruciais para compensar os riscos significativos inerentes ao uso de filtros imperfeitos por grandes empresas de mídia social para remover conteúdo. No entanto, os mecanismos de apelação não parecem estar disponíveis para ações de fiscalização adotadas por essas empresas, como rotulagem e redução de alcance. Também não parecem estar disponíveis para contestar decisões baseadas em políticas de dano coordenado ou comportamento inautêntico. Além disso, não está claro se os mecanismos de apelação estão acessíveis em diversos idiomas."

[474] UNIÃO EUROPEIA. *2022 Strengthened Code of Practice on Disinformation.* Disponível em: <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>. Acesso em 15 fev. 2025.

Digital Service Act, reforçando seu papel dentro do arcabouço regulatório da União Europeia, com vigência a partir de 1º de julho de 2025.[475-476]

A nova versão do Código de Conduta possui 44 compromissos e 128 medidas que abrangem diversas áreas. No capítulo de "empoderamentos do usuário", existem propostas de normas procedimentais que visam garantir:

(i) *Denúncias*: canal aberto para usuários denunciarem informações falsas ou enganosas, garantindo a adoção de medidas contra abusos no uso das ferramentas de denúncia, como denúncias em massa mal-intencionadas;[477] e

(ii) *Notificações, Recursos e Relatórios de transparência*: notificar os usuários sobre ações de moderação tomadas contra seu conteúdo ou conta e o funcionamento do sistema de recurso, oferecendo um mecanismo transparente de apelação para revisão das decisões de moderação, com a divulgação do trâmite do processo ao usuário,

---

[475] O *Digital Services Act* (DSA) é uma legislação da União Europeia adotada em outubro de 2022, que visa criar um ambiente digital mais seguro, transparente e responsável. Seu objetivo é definir regras claras para serviços intermediários *online*, como redes sociais e plataformas digitais, protegendo os direitos fundamentais dos usuários e estabelecendo responsabilidades para as empresas do setor. Entre os principais pontos abordados pelo DSA estão a remoção de conteúdo ilegal, a transparência na publicidade e a regulação da disseminação de desinformação (COMISSÃO EUROPEIA. *The Digital Services Act*. Disponível em: <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en>. Disponível em: 15 fev. 2025).

[476] É importante avaliar que a integração do Código no regime jurídico da União Europeia pouco depois da manifestação de Mark Zuckerberg, CEO da Meta, sobre a remoção de checadores de fato, é uma questão marcante, cuja resposta foi imediata.

[477] UNIÃO EUROPEIA. *2022 Strengthened Code of Practice on Disinformation*…, p. 31: "*Funcionalidade para sinalizar informações falsas e/ou enganosas prejudiciais*. Compromisso 23. Os signatários relevantes comprometem-se a fornecer aos usuários uma funcionalidade para sinalizar informações falsas e/ou enganosas prejudiciais que violem as políticas ou os termos de serviço dos signatários. Para cumprir o Compromisso 23: Medida 23.1. Os signatários relevantes desenvolverão ou continuarão a disponibilizar, em todos os seus serviços e em todos os idiomas dos Estados-Membros em que operam, uma funcionalidade de fácil uso para que os usuários possam sinalizar informações falsas e/ou enganosas prejudiciais que violem suas políticas ou termos de serviço. Essa funcionalidade deverá resultar em ações de acompanhamento adequadas, proporcionais e consistentes, respeitando plenamente a liberdade de expressão. QRE 23.1.1: Os signatários relevantes relatarão a disponibilidade de sistemas de sinalização para suas políticas relacionadas a informações falsas e/ou enganosas prejudiciais em todos os Estados-Membros da UE e especificarão as diferentes etapas necessárias para acionar esses sistemas. Medida 23.2. Os signatários relevantes adotarão as medidas necessárias para garantir que essa funcionalidade esteja devidamente protegida contra abusos humanos ou automatizados (por exemplo, a prática de "denúncias em massa" para silenciar outras vozes). QRE 23.2.1: Os signatários relevantes relatarão as medidas gerais adotadas para garantir a integridade de seus sistemas de denúncias e recursos, sem divulgar informações que possam ajudar potenciais abusadores a identificar e explorar vulnerabilidades nesses sistemas." (Tradução nossa).

além da publicação de dados sobre denúncias, ações de moderação e apelações no Centro de Transparência.[478]

3.3.3.8. Princípios de Manila

Conforme já antecipado acima, os Princípios de Manila[479] representam um marco fundamental na discussão sobre a responsabilidade legal dos intermediários da internet, como provedores de acesso, redes sociais e mecanismos de busca, em relação ao conteúdo gerado por terceiros. Os princípios foram publicados em 2015, após um processo colaborativo que envolveu especialistas em direitos humanos, tecnologia e políticas públicas.[480]

Os Princípios de Manila oferecem diretrizes essenciais para assegurar que o devido processo seja observado na restrição de conteúdo, protegendo tanto os interesses das partes envolvidas quanto a liberdade de expressão, seguindo uma série de normas procedimentais, como os Princípios III e V, que tratam do devido processo na requisição de restrição de conteúdo e na própria moderação.

De acordo com o Princípio III, as solicitações para restringir conteúdo devem ser claras, inequívocas e seguir o devido processo. Em síntese, as requisições devem atender a critérios mínimos, como: *(i)* a base legal que sustenta a alegação de ilegalidade; *(ii)* o identificador da internet (como URL) e uma descrição do

---

[478] *Idem*: "*Mecanismo de recurso transparente*. Compromisso 24. Os signatários relevantes comprometem-se a informar os usuários cujos conteúdos ou contas tenham sido alvo de ações de cumprimento (como rotulagem, redução de alcance ou outras medidas) devido a violações de políticas relevantes para esta seção (conforme descrito na Medida 18.2). Além disso, comprometem-se a fornecer aos usuários a possibilidade de recorrer dessas ações por meio de um mecanismo transparente, garantindo que as reclamações sejam tratadas de maneira ágil, diligente, transparente e objetiva, e revertendo a ação sem demora indevida quando a reclamação for considerada válida. Para cumprir o Compromisso 24: Medida 24.1. Os signatários relevantes comprometem-se a fornecer aos usuários informações sobre os motivos pelos quais determinado conteúdo ou conta foi rotulado, teve seu alcance reduzido ou foi alvo de outra ação de cumprimento, bem como a base para tal decisão e a possibilidade de recorrer por meio de um mecanismo transparente. QRE 24.1.1: Os signatários relevantes relatarão a disponibilidade de seus sistemas de notificação e apelação em todos os Estados-Membros e idiomas, detalhando as etapas do procedimento de recurso. SLI 24.1.1: Os signatários relevantes fornecerão informações sobre o número e a natureza das ações de cumprimento tomadas com base nas políticas descritas na Medida 18.2, o número de ações que foram objeto de recurso, os resultados desses recursos e, na medida do possível, métricas que ofereçam uma visão sobre a duração e a eficácia do processamento dos recursos. Essas informações serão publicadas no Centro de Transparência." (Tradução nossa).

[479] Os principais colaboradores do comitê que elaborou o documento são integrantes das organizações Electronic Frontier Foundation (EFF, USA), Center for Internet and Society (CIS, Índia), Article 19 (Reino Unido), KICTANET (Quênia), Derechos Digitales (Chile), Asociación por los Derechos Civiles (ADC, Argentina) e Open Net (Coreia do Sul).

[480] ELETRIC FRONTIER FOUNDATION et al. Princípios de Manila sobre Responsabilidade de Provedores. Disponível em: <https://manilaprinciples.org/index.html>. Acesso em 17 fev. 2025.

conteúdo supostamente ilegal; *(iii)* a consideração das limitações, exceções e defesas disponíveis ao provedor do conteúdo; (*iv*) os dados de contato de quem está fazendo a solicitação ou de seu representante, a menos que a lei proíba; *(v)* evidências que comprovem a legitimidade para fazer a requisição; *(vi)* uma declaração de boa-fé, afirmando que as informações fornecidas são precisas; e *(vii)* a indicação de quando o conteúdo for moderado mediante solicitação.[481]

Já o Princípio V trata especificamente do direito ao contraditório, à instância recursal e ao restabelecimento do conteúdo. Após a notificação de requisição de restrição de conteúdo, as plataformas que hospedam conteúdo devem *(i)* encaminhar a solicitação ao usuário ou notificar o reclamante sobre a impossibilidade de fazê-lo; *(ii)* fornecer uma explicação clara e acessível sobre os direitos do usuário, incluindo mecanismos de contranotificação e recurso; *(iii)* a plataforma e o provedor de conteúdo devem ter a oportunidade de se manifestar, exceto em circunstâncias excepcionais – *v.g.*, pornografia infantil, incitação ao genocídio ou terrorismo e discurso de ódio; *(iv)* a notificação deve indicar a penalidade de solicitações de má-fé ou repetidamente requeiram a restrição de conteúdo de forma infundada; e *(v)* manter a proteção dos dados pessoais dos usuários, com sigilo de sua identificação, e não requerer sua identificação sem autorização judicial.[482]

3.3.3.9. Princípios de Santa Clara

Em 2018, entidades de direitos humanos, ativistas e peritos acadêmicos desenvolveram os já mencionados Princípios de Santa Clara. Dentre suas colaborações, os Princípios de Santa Clara propõem deveres procedimentais (notificação, revisão e relatório) que podem auxiliar no aperfeiçoamento do regime jurídico brasileiro.[483] Os Princípios de Santa Clara estabelecem, essencialmente, os seguintes deveres procedimentais:

---

[481] *Ibidem*, pp. 30-35.
[482] *Ibidem*, pp. 40-48.
[483] ACCESS NOW *et al*. *Os Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo*. Disponível em: <https://santaclaraprinciples.org/>. Acesso em 12 jan. 2025.

(i) *Notificações*: as plataformas devem informar os usuários sobre remoção de conteúdo, suspensão de conta ou outras ações tomadas por violação das regras e políticas, com exceção para os casos de *spam*, *phishing* ou *malware*, devendo essa notificação conter: *(a)* URL, trecho do conteúdo ou outra informação suficiente para identificação; *(b)* cláusula específica da diretriz ou regra violada; *(c)* método de detecção (denúncia de usuário, revisão confiável, automação ou determinação legal); *(d)* informação sobre envolvimento governamental e, se aplicável, referência à legislação violada; *(e)* notificação tempestiva e explicação clara do processo de apelação; *(f)* disponibilidade da notificação mesmo após suspensão da conta, *(g)* acesso ao histórico de moderação para usuários que reportam conteúdo; *(h)* redação na língua original do conteúdo ou na interface do usuário; e *(i)* informações sobre canais de suporte;

(ii) *Recursos*: os usuários devem ter acesso a recursos e processos de revisão para contestar a remoção de conteúdo, suspensão de conta ou outras sanções, devendo ser garantindo ao usuário: *(a)* um mecanismo claro e acessível para envio de apelação; *(b)* informações sobre a existência de revisão independente, se houver; *(c)* revisão por equipe diferente da que tomou a decisão inicial; *(d)* avaliação feita por pessoas com conhecimento sobre a língua e contexto do conteúdo; *(e)* oportunidade para o usuário apresentar informações adicionais; e *(f)* notificação do resultado da revisão com justificativa suficiente;

(iii) *Relatórios de transparência:* as plataformas digitais devem divulgar dados sobre a moderação de conteúdo para garantir transparência por meio de relatórios de transparência, com divulgação em formato aberto e acessível, que devem conter: *(a)* número total de conteúdos removidos e contas suspensas; *(ii)* número de apelações e suas taxas de êxito; *(iii)* número de conteúdos restaurados proativamente sem apelação; *(iv)* dados sobre remoções relacionadas a discurso de ódio, crises (como pandemias e conflitos) e pedidos de governos; *(v)* número e origem dos pedidos governamentais para remoção de conteúdo; *(vi)* base legal ou regras empresariais que justificaram a

> remoção; *(vii)* como e quando a automação é usada; *(viii)* tipos de conteúdo moderados automaticamente; *(ix)* critérios e taxas de precisão desses processos; *(x)* existência de supervisão humana; *(xi)* número de apelações exitosas sobre conteúdo removido por automação; e *(xii)* participação em bancos de dados de *hashs* compartilhados.

Essas recomendações consistem, portanto, em um "um conjunto de critérios de referência ou primeiros passos que as empresas envolvidas na atividade de moderação de conteúdo" devem adotar para "oferecer um devido processo significativo para os usuários impactados e para garantir que a aplicação de suas diretrizes de conteúdo é justa, sem viés, proporcional e que respeita os direitos dos usuários".[484]

3.3.3.10. Global Partners Digital (GPD)

Em dezembro de 2017, a GPD respondeu à consulta do Relator Especial da ONU sobre Liberdade de Expressão acerca da regulação das plataformas, redes sociais e buscadores.[485] Em análise das normas procedimentais, verifica-se que a GPD aponta a necessidade de a plataforma:[486]

> *(i)* *Notificação*: enviar uma notificação ao usuário afetado pela moderação sobre o conteúdo restrito, removido ou suspenso ou sobre sanções à sua conta na rede social, devendo ser consideradas possíveis barreiras linguísticas;
>
> *(ii)* *Recursos*: explicar o processo de apelação de forma clara, com prazos indicativos para cada etapa, como as decisões são tomadas e os possíveis resultados, garantindo que os usuários tenham acesso razoável a informações, conselhos e expertise para participar efetivamente do processo de apelação;

---

[484] *Idem*.
[485] GLOBAL PARTNERS DIGITAL. *Content Regulation in the Digital Age*. OHCHR, sem data. Disponível em: <GlobalPartnersDigital.pdf>. Acesso em 14 fev. 2025.
[486] *Ibidem*, pp. 14-18.

(iii) *Relatórios de transparência*: incluir relatórios sobre o número de apelações, taxas de êxito e estudos de caso, com a transparência dos padrões aplicados, sua interpretação e os processos de decisão. Isso inclui publicar orientações sobre os termos de serviço, detalhes sobre o uso de mecanismos automatizados e estatísticas regulares sobre solicitações de remoção de conteúdo e seus resultados; e

(iv) *Mecanismos automatizados*: garantir que o uso de automação ou algoritmos para regular conteúdo seja acompanhado por supervisão humana para a revisão das decisões automatizadas.

Em 15 de maio de 2018, a GPD explicou que a moderação de conteúdo tem ameaçado o exercício legítimo da liberdade de expressão por alguns regimes jurídicos, como o alemão (*Netzdg*), que dispõe a necessidade de moderação de conteúdo em menos de 24 horas para plataformas com mais de 2 milhões de usuários, sob pena de aplicação de multas de até €50 milhões de euros.

O Comissário de Assuntos Internos da EU sugeriu o prazo de duas horas para a remoção de conteúdo. No Reino Unido, o Ministro do Interior também sugeriu a aplicação de multas às plataformas que falharem em moderar conteúdo considerado “radical” ou “extremista”. Por causa disso, a GPD propôs uma moderação de conteúdo focada em normas procedimentais, com a aplicação de deveres de transparências, bem como a criação de uma entidade internacional e independente que preste consulta às grandes plataformas sobre as melhores condutas e/ou interpretações das normas e diretrizes internacionais.[487]

3.3.3.11. Conclusão preliminar

Com base no estudo de propostas apresentadas por organizações nacionais e internacionais e na análise de pesquisadores que se debruçaram sobre o tema, é possível constatar o desenvolvimento de um padrão de orientação normativa e de

[487] GLOBAL PARTNERS DIGITAL. Disponível em: <Content regulation laws threaten our freedom of expression. We need a new approach – Global Partners Digital>. Acesso em 14 fev. 2025.

boas práticas no contexto da moderação de conteúdo congruente com o respeito às garantias de transparência, devido processo e isonomia.

Em vista disso, buscou-se esquematizar abaixo um padrão de deveres procedimentais, baseados nas propostas acima analisadas e nos ensinamentos doutrinários sobre tema, que poderiam corroborar com a proteção dos direitos humanos e assegurar o legítimo exercício da liberdade de expressão dos usuários nas redes sociais, conforme sistematizado a seguir:

(i) *Denúncias:* as solicitações/denúncias com pedido de remoção, suspensão ou despromoção de conta e/ou conteúdo devem ser feitas por meio de *(a.1)* um sistema simplificado de denúncia, devendo ser indicado *(a.2)* o identificador da internet (como URL) e uma descrição do conteúdo supostamente ilegal, *(a.3)* o fundamento dos termos de uso ou do ordenamento jurídico que foi supostamente violado pela conta e/ou conteúdo, e acompanhado de *(a.4)* uma declaração de que as informações são precisas e verdadeiras. Nesse contexto, os denunciantes deverão ser comunicados sobre a possibilidade de aplicação de penalidades em caso de solicitação de má-fé ou solicitações repetitivas sem fundamento jurídico adequado;

(ii) *Notificações:* as notificações aos usuários sobre as decisões de moderação, que deverão ser excepcionadas em casos de graves, a serem definidos por um órgão fiscalizador/regulador independente, como, por exemplo, incitação à violência, pornografia infantil, terrorismo, suicídio etc., devem conter: *(ii.1)* a identificação da conta e/ou conteúdo, *(ii.2)* o fundamento da violação às regras da plataforma, *(ii.3)* a explicação do procedimento de moderação de conteúdo, *(ii.4)* a origem do requerimento de restrição ou remoção de conteúdo, *(ii.5)* o prazo da contranotificação, com possibilidade de apresentar explicações e/ou documentos adicionais, *(ii.6)* o prazo recursal para a revisão do entendimento por revisor que não originou a moderação contestada;

(iii) *Recursos*: deve ser assegurado o direito à apresentação de recurso *(iv.1)* aos usuários que tenham contas ou conteúdos afetados pelas

decisões de moderação de conteúdo e *(iv.2)* aos denunciantes que utilizarem o sistema de denúncia (*flagging*) e que tiverem sido diretamente afetados pelo conteúdo supostamente ilegal não removido pela plataforma na decisão de moderação, devendo ser garantido *(iv.3)* o acompanhamento do procedimento recursal e do resultado final do recurso pelas partes interessadas e que (*iv.4*) o recurso seja julgado *(a)* de forma motivada, com o enfrentamento das razões recursais e a indicação das regras das plataformas suscitadas pelo recorrente e aplicáveis ao caso, *(b)* dentro de prazo razoável, *(c)* preferencialmente por revisores humanos não envolvidos no julgamento inicial; e *(iv.5)* em caso de provimento do recurso, ser restabelecido o conteúdo ou a conta objeto da decisão de moderação tomada pela plataforma;

(iv) *Publicação dos termos de uso e regras de comunidade*: a plataforma deverá publicar os termos de uso e das regras de comunidade e das diretrizes internas que norteiam os moderadores humanos na tomada de decisões nos casos concretos, para que os usuários compreendam o racional das regras estabelecidas e sua aplicação prática, bem como informar de forma clara e tempestiva de alterações nos termos de uso, regras de comunidade ou nas diretrizes internas de moderação. É fundamental que a plataforma informe ainda aos usuários e ao público em geral sobre o funcionamento dos mecanismos de moderação automatizada ou humano, como, por exemplo, em quais casos específicos são utilizados os mecanismos automatizados (algoritmos) e/ou é feita a revisão por moderadores humanos;

(v) *Relatórios de Transparência*: as plataformas deverão publicar relatórios de transparência periódicos, em formato aberto, acessível e com informações uniformizadas e padronizadas entre si, contendo o seguinte: *(v.1)* número total de conteúdos removidos e contas suspensas; *(v.2)* número de apelações e suas taxas de êxito; *(v.3)* número de conteúdos restaurados proativamente sem apelação; *(v.4)* dados sobre remoções relacionadas a discurso de ódio, crises (como pandemias e conflitos) e pedidos

> governamentais; *(d.5)* número e origem dos pedidos governamentais para remoção de conteúdo; *(v.6)* base legal ou regras empresariais que justificaram a remoção; *(v.7)* como e quando a automação é usada; *(v.8)* tipos de conteúdo moderados automaticamente; *(d.9)* critérios e taxas de precisão desses processos; *(v.10)* os casos solucionados por supervisão humana; *(v.11)* número de apelações exitosas sobre conteúdo removido por automação; e *(v.12)* informações detalhadas sobre os bancos de dados de *hashes* compartilhados.

As sugestões acima esquematizadas para o estabelecimento de obrigações de notificação, recursos e relatórios de transparência, decorrentes das conclusões da pesquisa desenvolvida nesta dissertação, visam a contribuir para o cumprimento dos deveres procedimentais de transparência, devido processo e isonomia pelas plataformas digitais.

As obrigações de aprimoramento dos mecanismos de denúncias e notificações (itens "i" e "ii" acima), já existentes em diversas plataformas, podem ser implementadas, primordialmente, por meio de mecanismos ou sistemas automatizados próprios das plataformas digitais, com supervisão humana, para garantir a acuracidade da conformidade com as regras propostas.

A exigência de um sistema recursal das decisões de moderação (item "iii" acima), embora gere custos agregados pela necessidade de contratação, treinamento e aperfeiçoamento de times de revisores humanos, justifica-se no imperativo da própria legitimidade decisória das plataformas digitais, tratando-se uma questão central da própria atividade de moderação de conteúdo e, por consequência, da própria atividade econômica das redes sociais.

Os deveres relacionados à divulgação de informações claras, objetivas e transparentes (itens "iv" e "v" acima), por meio da publicação de seus termos de uso, regras de comunidades e diretrizes de moderação, bem como de relatórios de transparência, são fundamentais para a higidez do procedimento de moderação de conteúdo, garantindo transparência e *accountability* das plataformas.

Entende-se que o estabelecimento desses deveres procedimentais mínimos não acarretará dispêndios desnecessários ou excessivamente onerosos aos negócios das plataformas digitais. O custo de conformidade com essas novas regras

procedimentais, inicialmente, será mais elevado, mas tendente à redução, uma vez que essas atividades sejam incorporadas à própria atividade das plataformas.

De todo modo, essas obrigações somente devem ser exigidas das grandes plataformas transnacionais, com milhões de usuários em território nacional – *v.g.*, o *Facebook*, *X* e *Youtube* –,[488] que possuem ou, ao menos, deveriam possuir uma estrutura organizacional capaz de implementar essas regras procedimentais. Essa abordagem visa evitar efeitos anticoncorrenciais da imposição de obrigações que poderiam prejudicar plataformas digitais de menor porte.

Por meio dessas recomendações, busca-se garantir a proteção às garantias fundamentais dos usuários, em especial da liberdade de expressão, privacidade, honra, segurança, proteção de dados e de direitos autorais etc., sem causar, por outro lado, interferência excessiva sobre a atividade de moderação de conteúdo das redes sociais, em respeito à autonomia da vontade e à liberdade editorial das plataformas.

3.3.4. Órgão fiscalizador independente

Conforme debatido no subcapítulo 3.1 *supra*, o modelo de regulação puramente estatal gera riscos de uma excessiva intervenção sobre a liberdade de expressão no ambiente digital, trazendo à tona a indesejada censura. Não parece adequado, portanto, que um órgão constituído exclusivamente por representantes do Estado desempenhe o papel de fiscalizar ou regular o modelo de autorregulação regulada proposto. Para se evitar os possíveis riscos à liberdade de expressão no ambiente virtual, é recomendável que o papel de fiscalização e regulação das redes sociais seja desempenhado por um órgão independente e constituído majoritariamente por integrantes da sociedade civil.

Esse órgão – ainda amorfo – não teria o objetivo de regular todas as celeumas que surgirem individualmente entre as redes sociais e os usuários. Sua competência deve consistir em realizar uma "análise sistêmica do modelo instituído e administrado pelas plataformas e auditar os relatórios de transparência".[489]

---

[488] Por exemplo, o PL das Fake News previa a aplicação de normas e mecanismos de transparência para provedores de redes sociais, ferramentas de busca e de mensageria instantânea, com "número médio de usuários mensais no país superior a 10.000.000 (dez milhões)" (artigo 2º). (SILVA, Orlando. *Relatório final do Relator Deputado Orlando Silva*. Apresentado em 31 de março de 2022, no Plenário da Câmara dos Deputados. Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=PRLP+1+%3D%3E+PL+2630/2020>. Acesso em 15 fev. 2025).

[489] BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*..., p. 283.

Inclusive, algumas organizações já defenderam a compatibilidade entre a imunidade das redes sociais e a necessidade de aplicação de sanções administrativas por riscos sistêmicos (*v.g.*, ITS Rio).[490]

A questão central é saber *qual* entidade poderia desempenhar essa tarefa. O Comitê Gestor da Internet – CGI.br,[491] entidade sem personalidade jurídica ou *status* de órgão público, disponibilizou uma consulta pública com 43 perguntas com três eixos: 1) *quem regular*, 2) *o que regular* e 3) *como regular*. No questionamento n° 39, perguntou: "quais órgãos, agências ou autoridades públicas devem estar diretamente envolvidos com a implementação da regulação de plataformas digitais"?[492] Nas respostas aos questionamentos públicos do CGI.br, verificou-se o posicionamento majoritário de que não poderia haver a fiscalização e regulação concentrada apenas em um único órgão, especialmente entre aqueles que concordavam com a regulação das plataformas digitais.

---

[490] ITS Rio discute a compatibilidade entre a imunidade das plataformas digitais contra responsabilização por conteúdos individuais de terceiros, conforme o art. 19 do Marco Civil da Internet, e a possibilidade de aplicação de sanções administrativas por descumprimento do dever de cuidado frente a riscos sistêmicos, ainda em definição nas iniciativas regulatórias atuais (ITS RIO. Regulação de Plataformas Digitais, disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 16 fev. 2025). Conforme exposta na resposta: "(...) a análise da autoridade competente precisa se ater, por exemplo, às ações implementadas pela plataforma para endereçar um eventual risco sistêmico, deixando de fora qualquer juízo sobre conteúdos individuais de terceiros que permanecem regulados pelo art. 19 do Marco Civil da Internet. Em outras palavras, é possível compatibilizar a imunidade das plataformas da responsabilização civil por danos causados por conteúdos (individuais) de terceiros (nos termos do MCI) com a previsão da aplicação de sanções administrativas em razão do não cumprimento do chamado dever de cuidado diante de riscos sistêmicos, que ainda precisam ser definidos de maneira mais apurada nas presentes iniciativas regulatórias para que essas figuras não sejam abusadas no futuro." (ITS RIO. *Regulação de Plataformas Digital*. publicada por CGI.br. Disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 16 fev. 2025).

[491] SUNDFELD, Carlos Ari; ROSILHO, André. A governança não estatal da internet e o direito brasileiro. *RDA*, Rio de Janeiro, v. 270, pp. 41/79, set./dez. 2015. Disponível em: <https://periodicos.fgv.br/rda/article/view/58737/57530>. Acesso em 18 fev. 2025, aqui, p. 62: "Suas atribuições e objetivos, conforme consta do seu estatuto social, são múltiplas e diversificadas. A ele compete: 1) registrar nomes de domínio de primeiro nível; 2) distribuir endereços de IP; 3) operar computadores, servidores de rede e toda a infraestrutura necessária, de modo a garantir a boa funcionalidade da operação de registro e manutenção dos domínios sob o ".br"; 4) atender aos requisitos de segurança e emergência na internet brasileira em articulação e cooperação com as entidades e os órgãos responsáveis; 5) desenvolver projetos que visem melhorar a qualidade da internet no Brasil e disseminar seu uso; 6) fomentar e acompanhar a disponibilização e a universalização de serviços de internet no país; e 7) promover ou colaborar na realização de cursos, simpósios, seminários, conferências, feiras e congressos, visando contribuir para o desenvolvimento e o aperfeiçoamento do ensino e dos conhecimentos nas áreas de suas especialidades."

[492] CGI.br. *Consulta sobre Regulação de Plataformas Digitais*. Disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 18 fev. 2025.

A Anatel tentou convencer o Congresso Nacional a aumentar sua competência para abarcar a atuação sobre as plataformas digitais.[493] A Coalizão Direitos na Rede (CDR), no entanto, fez duras críticas em torno da proposta da agência, ao argumento de que ela não teria competência para fiscalizar e regular as redes sociais, por carecer de capacidade técnica e legitimidade democrática.[494]

Em análise das propostas disponíveis na consulta pública da CGI.br, entidades e pessoas físicas sugeriram a criação de uma entidade multisetorial que englobasse indicações legislativo, executivo e judiciário, membros do Ministério Público e da OAB, representantes de autarquias federais e órgãos públicos (*v.g.*, Conselho Administrativo de Defesa Econômica – CADE, Secretaria Nacional do Consumidor – SENACON, Autoridade Nacional de Proteção de Dados etc.).

Uma proposta promissora foi elaborada pela Comissão Especial de Direito Digital do Conselho Federal da OAB. Essa proposta propõe a criação do *Sistema Brasileiro de Regulação de Plataformas Digitais* (SBRPD), que busca regular a liberdade de expressão por um sistema de pesos e contrapesos.[495] A proposta considera a constituição de um sistema em três camadas:

(i) Conselho de Políticas Digitais (CPD), órgão deliberativo responsável por fiscalizar e aplicar as diretrizes legalmente estabelecidas, composto por membros indicados pelos três Poderes da República, além da indicação da Anatel, Cade, ANPD e OAB Federal;[496]

---

[493] BRANT, Danielle; GABRIEL, João. Anatel faz lobby para regular big techs e cogita criar estrutura contra fake news, publicado em 09 mai. 2023. Disponível em: <https://www1.folha.uol.com.br/poder/2023/05/anatel-faz-lobby-para-regular-big-techs-e-cogita-criar-estrutura-contra-fake-news.shtml>. Acesso em 16 fev. 2025.

[494] COALIZÃO DIREITOS NAS REDES. *Órgão independente de supervisão das plataformas é essencial, mas não pode ser Anatel*, publicado em 28 abr. 2023**.** Disponível em: <https://direitosnarede.org.br/2023/04/28/orgao-independente-de-supervisao-das-plataformas-e-essencial-mas-nao-pode-ser-anatel/?ref=nucleo.jor.br>. Acesso em 16 fev. 2025.

[495] COMISSÃO ESPECIAL DE DIREITO DIGITAL DO CONSELHO FEDERAL DA OAB. *Ofício nº 001/2023 – CEDD/OAB*, publicado em 13 mai. 2023. Disponível em: <https://nucleo.jor.br/content/files/2023/05/Ofi-cio-Sistema-brasileiro-de-regulac-a-o-de-plataformas-digitais.pdf>. Acesso em 16 fev. 2025.

[496] *Idem*: "Art. 53. O CPD é composto por nove membros, nomeados pelo Presidente da República, com mandato de 2 (dois) anos, admitida 1 (uma recondução), sendo: I - um membro indicado Câmara dos Deputados; II – um membro indicado pelo Senado Federal; III - um membro indicado pelo Supremo Tribunal Federal; IV - um membro indicado pelo Tribunal Superior Eleitoral; V - um membro indicado pelo Presidente da República, que presidirá o CPD; VI - um membro indicado pela Agência Nacional de Telecomunicações (ANATEL); VII - um membro indicado pela Autoridade Nacional de Proteção de Dados (ANPD); VIII - um membro indicado pelo Conselho

(ii) Comitê Gestor da Internet no Brasil (CGI.br), mantendo a competência de promover o debate sobre o tema no Brasil mediante a realização de estudos, recomendações e diretrizes;[497] e

(iii) uma ou mais entidades de autorregulação, pessoa jurídica de direito privado, com a responsabilidade de deliberar sobre casos concretos de moderação de conteúdo no âmbito das plataformas digitais.[498]

Segundo a proposta, esses órgãos deveriam fiscalizar o cumprimento de deveres procedimentais, além de assegurar que as plataformas adotem medidas para combater ou minimizar os impactos de conteúdos danosos. A supervisão independente também poderia equilibrar o custo-benefício inerente à moderação de conteúdo, como a necessidade de decisões rápidas *versus* a precisão e a ampliação

Administrativo de Defesa Econômica (CADE); e IX - um membro indicado do Conselho Federal da Ordem dos Advogados do Brasil (OAB)."

[497] *Idem*: "Art. 58. Serão atribuições do Comitê Gestor da Internet no Brasil (CGI.br), além daquelas previstas pelas Leis nº 12.965, de 23 de abril de 2014, e nº 13.853, de 8 de julho de 2019, as seguintes: I - realizar estudos, pareceres e propor diretrizes estratégicas sobre liberdade, responsabilidade e transparência na internet; II - realizar estudos e debates para aprofundar o entendimento sobre desinformação, e propor diretrizes para o seu combate, no contexto da internet e das redes sociais; III - apresentar diretrizes para a elaboração de código de conduta para os provedores de redes sociais, ferramentas de busca e mensageria instantânea, para a garantia dos princípios e objetivos estabelecidos nos arts. 3º e 4º; IV - realizar estudos sobre os procedimentos de moderação de contas e de conteúdos adotados pelos provedores de redes sociais, bem como sugerir diretrizes para sua implementação; V - fornecer diretrizes e subsídios para os termos de uso dos provedores de redes sociais e de serviços de mensageria instantânea; VI - publicar a relação dos provedores que se enquadram no disposto no art. 2o desta lei; VII - emitir diretrizes e critérios para a instauração dos protocolos de segurança de que trata esta Lei; e VIII - emitir diretrizes e requisitos para a análise de riscos sistêmicos de que trata esta Lei. Parágrafo único. Fica garantida a composição multisetorial do CGI.br para fins de cumprimento das suas competências, com participação do Poder Público, do setor empresarial, do terceiro setor e da comunidade técnico-científica."

[498] *Idem*: "Art. 61. O CPD reconhecerá pessoa jurídica de direito privado como entidade de autorregulação nos termos desta lei, se demonstrada: I - a capacidade de revisão de decisões de moderação de conteúdo e contas, a partir da provocação dos provedores ou dos afetados diretamente por uma decisão; II - a existência de órgão competente para tomar decisões, em tempo útil e eficaz, sobre a revisão de medidas de moderação adotadas pelos provedores; III - a independência e a especialidade de seus analistas; IV - a disponibilização de serviço eficiente de atendimento e encaminhamento de reclamações; V - a definição de requisitos claros, objetivos e acessíveis para a participação dos provedores de redes sociais e serviços de mensageria privada; VI - a inclusão, em seu quadro organizacional, de uma ouvidoria independente com a finalidade de receber, encaminhar e solucionar solicitações e críticas, inclusive por meio digital, e avaliar as atividades da entidade; § 1º O CPD poderá reconhecer mais de uma entidade de autorregulação, desde que verificados os requisitos desta lei, e se comprovada a necessidade de especialização temática. § 2º O reconhecimento pode ser revogado ou vinculado a requisitos suplementares se alguma das condições para o reconhecimento deixar de ser cumprida. § 3º O prazo de solução da solicitação deve ser de 5 (cinco) dias úteis, contados a partir do pedido de revisão do afetado pela decisão do provedor ou da provocação pelo provedor em situações de dúvidas sobre a legalidade do conteúdo. § 4º A entidade de autorregulação deverá emitir relatórios semestrais em atendimento ao disposto nesta Lei; § 5º A entidade de autorregulação aprovará resoluções de modo a regular seus procedimentos de análise. § 6º As decisões da entidade de autorregulação serão fundamentadas e públicas."

do direito de recurso *versus* a agilidade na resolução de conflitos. Nesse contexto, enquanto a entidade de autorregulação seria aquela que criará uma espécie de "jurisprudência", tornando as regras mais claras e consistentes,[499] o Conselho de Políticas Digitais é quem avaliaria os riscos sistêmicos da autorregulação das próprias plataformas, como também da própria entidade autorreguladora.

Embora existam aqueles que defendem que o papel de fiscalização deveria ser desempenhado pelo próprio CGI.br ,[500] a criação do SBRPD parece ser mais uma alternativa a ser considerada. O modelo proposto busca assegurar legitimidade decisória e maior confiabilidade à moderação de conteúdo das plataformas, reconhecendo a complexidade de aplicar regras globais em contexto local. A ideia de incluir representantes da sociedade civil no órgão de fiscalização independente responsável pela moderação de conteúdo na internet parece ser viável, podendo trazer benefícios, especialmente em um contexto em que a regulação de conteúdo *online* é um tema complexo e multifacetado.

Independente da instituição a ser escolhida pelo legislador, o papel do órgão fiscalizador independente deve compreender o estabelecimento de diretrizes de interesse público a serem adotadas pelas plataformas digitais (como as garantias procedimentais acima propostas), bem como a necessária supervisão da implementação e do cumprimento dessas diretrizes.

---

[499] A destituição dos checadores de fato da Meta demonstra uma grave fragilidade no regime de proteção dos usuários, que pode ser contornado caso exista uma agência reguladora que implemente as políticas definidas, de forma voluntária, pelas próprias empresas, impedindo mudanças drásticas no regime jurídico com a simples mudança de governos. Nesse sentido, MACCARTHY, Mark. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*…, p. 6: "(…) digital social media platforms are free to make promises to the public concerning their content moderation practices, or not, as they see fit. But they are not free to make promises to their users that they do not keep. The supervising regulatory agency would be authorized to enforce these promises as well as any disclosure obligations to ensure that the public, the regulators and researchers have sufficient information about how platforms' moderation practices and content-ordering techniques might exacerbate the distribution of problematic content." Tradução: "As plataformas de mídia social digital são livres para fazer ou não promessas ao público sobre suas práticas de moderação de conteúdo, conforme julgarem adequado. No entanto, não podem fazer promessas aos seus usuários sem cumpri-las. A agência reguladora responsável teria autoridade para fazer cumprir essas promessas, bem como quaisquer obrigações de transparência, garantindo que o público, os reguladores e os pesquisadores tenham informações suficientes sobre como as práticas de moderação e as técnicas de ordenação de conteúdo das plataformas podem agravar a distribuição de conteúdos problemáticos."

[500] Em sentido contrário, BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na Era Digital*..., p. 284: "Um órgão já existente que poderia desempenhar essa função no Brasil é o CGI.br (...). Considerando os anos de experiência do Comitê e a percepção compartilhada de sua efetividade, a concentração das funções nesse órgão já existente se apresenta como uma alternativa promissora."

As recomendações, decisões e diretrizes do órgão fiscalizador independente devem ser públicas, com justificativas claras e fundamentadas, podendo o órgão utilizar o mecanismo de consultas públicas para receber contribuições da sociedade para aperfeiçoamento do sistema de fiscalização e regulação

Para garantir a imparcialidade do órgão fiscalizador independente, parece fundamental assegurar sua autonomia financeira/orçamentária, protegendo suas decisões de pressões políticas e econômicas. Mas qual seria o modelo mais adequado para garantir essa independência? Seria uma estrutura semelhante à de uma autarquia, com orçamento próprio e autonomia administrativa, ou um modelo híbrido, que permitisse maior flexibilidade operacional?

Ainda há incertezas sobre a melhor forma de organizar o órgão fiscalizador independente, especialmente no que se refere à governança, critérios de nomeação e mecanismos de prestação de contas. A ausência de um desenho institucional bem definido pode comprometer sua eficácia e abrir margem para interferências indevidas. O debate sobre sua constituição precisa ser aprofundado, em pesquisa específica sobre o tema, equilibrando autonomia, transparência e previsibilidade.

3.3.5. Falhas sistêmicas no cumprimento de deveres procedimentais

No contexto do estabelecimento de deveres procedimentais à moderação de conteúdo das redes sociais, como transparência, devido processo e isonomia, deve-se ponderar a possibilidade de ocorrerem falhas pontuais, devido ao volume massivo de conteúdo publicado diariamente nas plataformas e à inevitabilidade de utilização de mecanismos automatizados na moderação de conteúdo.[501]

A melhor solução para assegurar o cumprimento dessas garantias procedimentais e evitar a ocorrência de remoção excessiva parece ser a

[501] *Ibidem*, p. 274: "Devido à enorme quantidade de conteúdo publicado nas plataformas e à inevitabilidade do uso de mecanismos automatizados para moderação de conteúdo, a responsabilização das plataformas por violação aos deveres de devido processo legal e da isonomia não deve ocorrer em casos pontuais e específicos, mas apenas quando a análise revele uma falha sistêmica em atender a esses deveres. Exatamente por isso, desde que sistemicamente observados esses requisitos de transparência, devido processo legal e isonomia, entende-se pela inadequação de decisões judiciais que determinem a restituição de conteúdo removido pelas plataformas com fundamento em termos e condições privados. Isso porque, desde que o façam observando deveres procedimentais, as plataformas devem ter liberdade de iniciativa e de expressão para definirem o tipo de plataforma que querem oferecer aos seus usuários."

responsabilização das plataformas digitais quando a análise revelar falhas sistêmicas, e não na ocorrência de falhas ou descumprimentos pontuais.

Evelyn Douek afirma que a natureza sistêmica das falhas é demonstrada quando, mediante a análise dos dados fornecidos durante relatórios de transparência e divulgação da forma de cumprimento de deveres procedimentais, como o devido processo e isonomia, os erros decorrem de uma escolha *ex ante* do sistema automatizado de moderação ou, ainda, na ausência de uma hierarquia ou protocolo revisional adequado.[502] Nesse sentido, a autora defende que:

> A revisão caso a caso é um modelo ineficaz de supervisão da moderação de conteúdo, pois tal abordagem, primeiro, falhará em identificar falhas sistêmicas e, segundo, distorcerá a tolerância ao risco ao destacar erros que podem ser o resultado de decisões razoáveis tomadas previamente no nível do sistema. Primeiro, uma abordagem baseada em postagens individuais 'não consegue sequer enxergar' danos agregados. O impacto desigual de um sistema não pode ser identificado ao se analisar uma única decisão, e isso é especialmente verdadeiro para a moderação automatizada. (...) Após anos descartando reclamações de viés, o Facebook anunciou que reformularia seus algoritmos de detecção de discurso de ódio para corrigir o impacto desigual de seus sistemas 'neutros em relação à raça'. O Perspective API, uma ferramenta utilizada por diversas plataformas, incluindo o Reddit e a segunda maior plataforma da América Latina, Taringa!, demonstrou marcar desproporcionalmente falas de usuários negros como discurso de ódio. A revisão individualista ex post tratará os resultados individuais desse viés como erros a serem corrigidos, em vez de reconhecer a necessidade de uma reforma sistêmica. Nem todas as falhas sistêmicas são resultado de uma IA defeituosa. Os erros podem ser consequência de outras formas de design problemático do sistema, que não serão evidentes na análise isolada de uma decisão. Um exemplo disso é o fato de o Facebook não ter removido um grupo paramilitar autodenominado Kenosha Guard, mesmo após este ter emitido um 'chamado às armas' antes de um protesto em Kenosha que resultou em mortes. Esse caso poderia ser enquadrado como uma falha individual do Facebook na aplicação de suas regras contra incitação à violência – um 'erro operacional', como Zuckerberg o chamou. Dessa perspectiva, o erro é um problema lamentável, mas inevitável, que deve ser corrigido posteriormente. Entretanto, a falha do Facebook em Kenosha foi uma falha de design do sistema. O evento do Kenosha Guard foi reportado mais de 455 vezes, representando a maioria esmagadora dos relatos de eventos daquele dia. Falhar na revisão adequada de um

[502] DOUEK, Evelyn. Content Moderation as Systems Thinking. *Harvard Law Review*, vol. 136, 2022, p. 43-44. Disponível em: <https://ssrn.com/abstract=4005326>. Acesso em 17 fev. 2025.

> item responsável pela maior parte dos relatórios diários não é um caso de escala inadministrável – é um colapso sistêmico.[503]

De fato, não é recomendável analisar isoladamente e sob uma ótica individualizada a moderação de conteúdo nas plataformas digitais, para determinar se um conteúdo específico viola as diretrizes estabelecidas. Essa abordagem fragmentada não leva em consideração o impacto sistêmico das decisões de moderação e a forma como elas podem afetar o ecossistema digital como um todo.

Portanto, é importante que a regulação e atuação do órgão fiscalizador independente vise a identificação e mitigação de *riscos sistêmicos*. Esse paradigma permite que reguladores e plataformas enfrentem os desafios da moderação com mais eficiência e adotem estratégias que reduzam os danos em larga escala. Para mitigar riscos sistêmicos, órgãos de fiscalização e regulação devem desenvolver mecanismos para monitorar padrões de comportamento e identificar tendências de falhas sistêmicas no ambiente digital.[504]

Os reguladores devem estabelecer, portanto, normas procedimentais para garantir que as plataformas adotem um dever de cuidado diante de riscos sistêmicos no contexto da moderação de conteúdo. Isso inclui exigências para que realizem avaliações regulares do impacto de suas práticas de moderação, garantindo que suas

---

[503] *Idem*: "Case-by-case review is a poor model of content moderation oversight because such review will, first, fail to identify systemic failures and, second, skew risk tolerance by highlighting mistakes that may be the product of reasonable ex ante decisions at the systems level. First, a post-by-post approach 'cannot even see' aggregate harms. A system's disparate impact cannot be identified by looking at a single decision, and this is especially true for automated moderation. (…) After years of dismissing complaints of bias, Facebook announced that it would overhaul its hate speech detection algorithms to address the disparate impact of its "race-blind" systems. Perspective API, a tool used by a wide variety of platforms including Reddit and Latin America's second-largest platform, Taringa!, has been shown to disproportionately flag black users' speech as hate speech. Ex post individualistic review will treat individual outcomes of such bias as errors to be corrected rather than potential evidence of the need for systemic reform. Not all systemic failures are the result of broken AI. Errors may be downstream from other forms of problematic system design that will not be apparent within the four corners of any individual decision. Facebook leaving up a self-proclaimed militia group called Kenosha Guard, after it issued a 'call to arms' in advance of a protest in Kenosha which turned deadly, is an example. This might be framed as an individual failure by Facebook to enforce its rules prohibiting incitement—an 'operational mistake,' as Zuckerberg called it. From this perspective, the error is a regrettable but inevitable product of operating at scale, to be corrected ex post. But Facebook's failure in Kenosha was a failure of system design. The Kenosha Guard's event was reported over 455 times, making up a staggering majority of event reports that day. Failing to properly review an item responsible for most of a day's reports is not a story of unmanageable scale—it is a systemic breakdown. Conversely, an error may be evidence of poor system design, but a single error may not be proof of system failure at all. Because platform scale means even the most carefully designed system will make mistakes, an error might be the consequence of a calculated and reasonable ex ante trade-off between differing values."

[504] *Ibidem*, pp. 77-78.

políticas não incentivem, ainda que indiretamente, a disseminação de conteúdos falsos ou danosos.

CONSIDERAÇÕES FINAIS

Ao longo da presente dissertação, buscou-se investigar as complexidades do atual sistema de moderação de conteúdo das redes sociais, analisando possíveis soluções regulatórias para mitigar o abuso do direito à liberdade de expressão no ambiente virtual.

Partindo da constatação de que a autorregulação das plataformas digitais, baseada na aplicação de seus termos de uso, apresenta limitações significativas, propõe-se um modelo de autorregulação regulada, também conhecida como corregulação, como solução mais adequada para o contexto brasileiro.

O estudo percorreu uma trajetória multidisciplinar, analisando diferentes perspectivas jurídicas, sociológicas e tecnológicas. Iniciou-se com um aprofundamento histórico da legislação brasileira sobre a liberdade de expressão, destacando as contradições entre a garantia constitucional deste direito e a prática reguladora, marcada por períodos de censura e intervenção estatal.

A análise da legislação, desde o início do século XIX até a edição do Marco Civil da Internet (MCI), evidenciou o movimento pendular entre a defesa intransigente da liberdade e a imposição de limites à liberdade de expressão para proteção de outros direitos fundamentais ou do próprio ordenamento jurídico.

A perspectiva histórica foi fundamental para contextualizar o debate sobre a moderação de conteúdo, especialmente em face do avanço tecnológico, que ampliou exponencialmente a capacidade de propagação de informações, tanto lícitas quanto ilícitas. A era digital, caracterizada pela descentralização do discurso público, impôs desafios sem precedentes, incluindo a disseminação de *fake news*, discursos de ódio e conteúdos danosos nas redes sociais.

A análise crítica do MCI revelou suas lacunas e limitações. Embora o MCI assegure a liberdade de expressão e proíba a censura prévia, admitindo a responsabilidade civil das plataformas apenas em caso de descumprimento de ordem judicial específica, a norma de regência não endereçou o vácuo normativo a respeito da moderação de conteúdo nas redes sociais,[505] limitando o papel do

[505] Há de se ressalvar, no entanto, que, à época da tramitação do MCI nas casas legislativas, nos idos dos anos 2010, o contexto da moderação de contéudo no Brasil era distinto, até mesmo porque o fenômeno das *fake news* ainda não havia sido pautado como um dos principais problemas da sociedade contemporânea, com relevantes impactos sobre temas caros à democracia, à saúde e segurança pública etc. Ao ser argumentar que existe um "vácuo normativo" no MCI, considera-se, portanto, o atual contexto nacional e internacional de debates a respeito da moderação de conteúdo.

Judiciário. Essas limitações são alvo de diversas críticas, sendo considerada por alguns autores como um retrocesso no regime de responsabilização das plataformas.

A evolução dos julgamentos do STF sobre a constitucionalidade do artigo 19 do MCI ilustrou as controvérsias sobre a responsabilização e as normas procedimentais mínimas que deveriam adotadas pelas plataformas digitais, revelando a complexidade e atualidade do debate sobre a necessidade de regulação no contexto brasileiro.

O estudo apresentou uma análise aprofundada dos mecanismos de moderação de conteúdo utilizados pelas principais plataformas digitais (*Facebook*, *Youtube* e *X*), envolvendo a combinação de ferramentas automatizadas (algoritmos) e a revisão humana, com foco em suas políticas internas de moderação, para demonstrar os desafios práticos da moderação massiva de conteúdos e as falhas sistêmicas cometidas pelas plataformas no contexto dessa moderação.

No curso dessa análise, destacou-se a necessidade de implementação de mecanismos de transparência e *accountability* para garantia da legitimidade decisória das redes sociais, bem como os principais desafios práticos da autorregulação, incluindo: opacidade algorítmica, falta de consistência na aplicação das regras, vieses discriminatórios e a remoção excessiva de conteúdo.

A análise dos diferentes modelos de regulação estrangeira (CDA, *NetzDG*, *Online Safety Bill*, DSA) e dos documentos de *soft law* (Princípios de Manila e Princípios de Santa Clara) permitiu compreender as melhores práticas previstas na legislação de outros países e defendidas por importantes organizações internacionais, contribuindo, assim, para o desenvolvimento de uma proposta de corregulação das redes sociais para o Brasil.

Os modelos de regulação analisados apontaram para a necessidade de uma atuação conjunta do Estado e dos agentes privados, com o objetivo de combinar a expertise técnica das plataformas com o dever de proteção de interesses públicos. Essa visão de corregulação ou autorregulação regulada destaca a necessidade de criar um arcabouço normativo que inclua normas procedimentais, que garantam transparência, devido processo e isonomia na moderação de conteúdo.

A proposta de autorregulação regulada visa a construção de um modelo *policêntrico* que equilibra a atuação do Estado e das plataformas digitais na construção da solução regulatória. A proposta sugere uma atuação estatal que fomente a criação de um ambiente digital saudável, assegurando a liberdade de

expressão e outros direitos fundamentais dos usuários, sem sufocar a inovação e a liberdade editorial das plataformas.

A participação ativa dos agentes regulados, em especial das plataformas digitais, na formulação e implementação da regulação é essencial para a construção de um modelo transparente, legítimo e eficaz que atenda às peculiaridades do contexto brasileiro. Esse modelo não apenas contribui para a proteção dos direitos humanos, mas também para a segurança do sistema democrático, o que está diretamente relacionado à participação e conscientização do usuário, além da *accountability* e transparência no procedimento de moderação de conteúdo.

O estudo se propôs também a identificar um padrão nas orientações normativas e de boas práticas em torno da moderação de conteúdo, a partir do estudo de um grupo selecionado de organizações nacionais e internacionais. A partir dessa análise, observou-se recomendações convergentes sobre o estabelecimento de deveres procedimentais que podem ensejar o estabelecimento de princípios fundamentais para o contexto regulatório brasileiro, como a transparência, o devido processo, a isonomia e a implementação de mecanismos de fiscalização (vide conclusão preliminar – subcapítulo 3.3.3.11 *supra*).

Em termos de transparência, uma exigência constante nas propostas das organizações nacionais e internacionais a respeito da moderação de conteúdo, inclui a necessidade de divulgação aos usuários, público em geral e a pesquisadores independentes, inclusive por meio de relatórios de transparência, informações sobre as regras de moderação, os métodos utilizados (incluindo uso de mecanismos automatizados), dados sobre a aplicação das regras e os resultados obtidos.

Outro aspecto fundamental extraído dessas orientações é a garantia do devido processo para os usuários afetados pelas decisões de moderação de conteúdo, do qual deve, necessariamente, emergir o direito de ser ouvido antes da tomada de qualquer medida ou sanção contratual, assim como o direito a recorrer contra decisões de moderação tomadas pelas plataformas.

Além da garantia do devido processo, as decisões de moderação de conteúdo devem garantir, salvo em hipóteses excepcionais, o tratamento igualitário entre todos os usuários, desviando-se de vieses discriminatórios na aplicação das regras de moderação de conteúdo pelas plataformas digitais, garantindo-se, assim, um processo justo e imparcial.

Um elemento crucial e complementar dessa proposta é a necessidade de serem implementados mecanismos independentes de fiscalização, para monitorar a atuação das plataformas digitais e garantir o cumprimento dos padrões de transparência e dos demais deveres procedimentais acima tratados, como o devido processo e a isonomia.

Daí, portanto, a relevância da criação de um órgão fiscalizador independente, com composição multissetorial que inclui representantes do Estado, da sociedade civil, da academia e das empresas de tecnologia. Esse órgão seria responsável pela elaboração de diretrizes, normas e códigos de conduta, bem como pela supervisão da implementação e do cumprimento da regulação, garantindo maior transparência e *accountability* das plataformas. Sua atuação deve estar focada na mitigação de riscos sistêmicos na moderação de conteúdo, incluindo a prevenção de vieses discriminatórios, remoção excessiva e censura colateral.

A presente dissertação buscou, portanto, contribuir para o avanço do debate sobre a regulação da moderação de conteúdo em redes sociais no Brasil, apresentando um modelo de autorregulação regulada que considera as complexidades do ambiente digital, promovendo os valores fundamentais da democracia e da liberdade de expressão.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ACCESSNOW *et al*. *Os Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo*. Disponível em: <https://santaclaraprinciples.org/>. Acesso em 12 jan. 2025.
2. AJZENMAN, Nicolás; CAVALCANTI, Tiago; DA MATA, Daniel. More than words: leaders' speech and risky behavior during a pandemic. *Cambridge-INET Working Paper Series*, nº 2020/19; Cambridge Working Papers in Economics: 2034, 2020. Disponível em: <https://econpapers.repec.org/paper/camcamdae/2034.htm>. Acesso em 28 fev. 2025.
3. ALLCOTT, Hunt; GENTZKOW, Matthew (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, vol. 31, n. 2, 2017, pp. 211–236. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211>. Acesso em 29 nov. 2024.
4. AMARAL, Francisco. *Direito civil: introdução*. 8ª ed. Rio de Janeiro: Renovar, 2014.
5. ANJOS, Alise Silva Martins; CASAM, Priscila Carla; MAIA, Janize Silva. As fake news e seus impactos na saúde da sociedade. *Pub Saúde*, 5, a141, 2021.
6. ARMIJO, Enrique. *Speech Regulation by Algorithm*. 30 Wm. & Mary Bill Rts. J. 245 (2021), pp. 245-263. Disponível: <https://scholarship.law.wm.edu/wmborj/vol30/iss2/3>. Acesso em 27 dez. 2024.
7. BALKIN, Jack M. How to Regulate (and Not Regulate) Social Media (November 8, 2019). *Journal of Free Speech Law* 71 (2021), Knight Institute Occasional Paper Series, No. 1 (March 25, 2020), Yale Law School, Public Law Research Paper Forthcoming, pp. 72-73. Disponível em: <https://ssrn.com/abstract=3484114>. Acesso em 29 nov. 2024.
8. BALKIN, Jack M., Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation (September 9, 2017). *UC Davis Law Review*, (2018 Forthcoming), Yale Law School, Public Law Research Paper No. 615, p. 3. Disponível em: <https://ssrn.com/abstract=3038939>. Acesso em 24 nov. 2024.
9. BARROSO, Luis Roberto. BARROSO, Luna Van Brussel. Prefácio à 3ª Edição. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021.
10. BARROSO, Luís Roberto. *Temas de direito constitucional*. Rio de Janeiro: Renovar, 2003.
11. BARROSO, Luís Roberto. Neoconstitucionalismo e Constitucionalização do Direito: o triunfo tardio do Direito Constitucional no Brasil. *Revista da Escola de Magistratura do Estado do Rio de Janeiro - EMERJ*, v. 9, nº 33, 2006, p. 43-92, p. 61. Disponível em: <https://www.emerj.tjrj.jus.br/revistaemerj_online/edicoes/revista33/Revista33_43.pdf>. Acesso em 5 jan. 2025.
12. BARROSO, Luis Roberto. Prefácio. In: Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*. Belo Horizonte: Fórum, 2022.

13. BARROSO, Luís Roberto; BARROSO, Luna van Brussel. Democracia, mídias sociais e liberdade de expressão: ódio, mentiras e a busca da verdade possível. *Direitos Fundamentais & Justiça*, Belo Horizonte, ano 17, n. 49, p. 285-311, jul./dez. 2023.
14. BARROSO, Luna van Brussel. *Liberdade de expressão e democracia na era digital: o impacto das mídias sociais no mundo contemporâneo*. Belo Horizonte: Fórum, 2022.
15. BBC. *China's WeChat, Weibo and Baidu under investigation*. BBC News, publicado em 11 ago. 2017. Disponível em: <https://www.bbc.com/news/world-asia-china-40896235>. Acesso em 13 jan. 2025.
16. BERMUDES, Sergio. *Introdução ao Processo Civil*. 4. ed. Rio de Janeiro: Forense, 2006.
17. BINENBOJM, Gustavo. Meios de comunicação de massa, pluralismo e democracia deliberativa. As liberdades de expressão e de imprensa nos Estados Unidos e no Brasil. *Revista da EMERJ*, v. 6, n. 23, 2003.
18. BINICHESKI, Paulo Roberto. *Responsabilidade Civil dos Provedores de Internet*. Curitiba: Juruá, 2011.
19. BOBBIO, Noberto; MATTEUCI, Nicola; PASQUINO, Gianfranco. *Dicionário político*. Vol. 1. 11ª ed. Brasília: Universidade de Brasília, 1998. (ebook).
20. BONAVIDES, Paulo. *Curso de Direito Constitucional*. 15ª ed. São Paulo: Malheiros, 2004.
21. BONIN, Robson. *Lira diz que PL das Fake News está morto e anuncia grupo por novo texto*. Veja, publicado em 09 mai. 2024. Disponível em: <https://veja.abril.com.br/coluna/radar/lira-diz-que-pl-das-fake-news-esta-morto-e-anuncia-grupo-por-novo-texto>. Acesso em 27 dez. 2024.
22. BRANT, Danielle; GABRIEL, João. *Anatel faz lobby para regular big techs e cogita criar estrutura contra fake new. UOL*, publicado em 09 mai. 2023. Disponível em: <https://www1.folha.uol.com.br/poder/2023/05/anatel-faz-lobby-para-regular-big-techs-e-cogita-criar-estrutura-contra-fake-news.shtml>. Acesso em 16 fev. 2025.
23. BRASIL. Advocacia-Geral da União (AGU*). Carta para resposta à notificação extrajudicial*. Disponível em: <https://www.gov.br/agu/pt-br/nota-agu-recebe-manifestacao-da-meta/Cartapararespostaanotificacaoextrajudicial_13.1.20251.pdf>. Acesso em 17 jan. 2025.
24. BRASIL. Câmara dos Deputados. *50 anos do Golpe de 1964*. <https://www2.camara.leg.br/atividade-legislativa/plenario/discursos/escrevendohistoria/destaque-de-materias/golpe-de-1964>. Acesso em 07 mar. 2025.
25. BRASIL. *Constituição da República dos Estados Unidos do Brasil (16 de julho de 1934)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao34.htm>. Acesso em 07 mar. 2025.
26. BRASIL. *Constituição da República dos Estados Unidos do Brasil (24 de fevereiro de 1891)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao91.htm>. Acesso em 07 mar. 2025.

27. BRASIL. Constituição da República Federativa do Brasil (24 de janeiro de 1967). Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao67.htm>. Acesso em 07 mar. 2025.
28. BRASIL. *Constituição dos Estados Unidos do Brasil (10 de novembro de 1937)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao37.htm>. Acesso em 07 mar. 2025.
29. BRASIL. *Constituição dos Estados Unidos do Brasil (18 de setembro de 1946)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao46.htm>. Acesso em 07 mar. 2025.
30. BRASIL. *Constituição politica do Imperio do Brazil (de 25 de março de 1824)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao24.htm>. Acesso em 07 mar. 2025.
31. BRASIL. *Decreto 592 – Pacto Internacional dos Direitos Civis e Políticos (6 de julho de 1992)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/decreto/1990-1994/d0592.htm>. Acesso em 07 mar. 2025.
32. BRASIL. *Decreto nº 678 – Pacto de San José da Costa Rica (6 de novembro de 1992)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/decreto/d0678.htm>. Acesso em 07 mar. 2025.
33. BRASIL. *Emenda Constitucional nº 1 (17 de outubro de 1969)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc_anterior1988/emc01-69.htm>. Acesso em 07 mar. 2025.
34. BRASIL. *Lei 5.250 (9 de fevereiro de 1967)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l5250.htm>. Acesso em 07 mar. 2025.
35. BRASIL. *Lei nº 12.695 – Marco Civil da Internet (23 de abril de 2014)*. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>. Acesso em 07 mar. 2025.
36. BRASIL. *Projeto de Lei 2.630*. Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=Tramitacao-PL%202630/2020>. Acesso em 07 mar. 2025.
37. BRASIL. Superior Tribunal de Justiça. 3ª Turma. Recurso Especial 2139749/SP. Rel. Min. Ricardo Villas Bôas Cueva. j. 27 dez. 2024.
38. BRASIL. Superior Tribunal de Justiça. AgInt nos EDcl no REsp n. 1.402.112/SE, Min. Rel. Lázaro Guimarães, 4ª Turma, j. 19 jun. 2018.
39. BRASIL. Superior Tribunal de Justiça. REsp n. 1.337.990/SP, Min. Rel. Paulo de Tarso Sanseverino, 3ª Turma, j. 21 ago. 2014.
40. BRASIL. Superior Tribunal de Justiça. REsp n. 1.568.935/RJ, Min. Rel. Ricardo Villas Bôas Cueva, Presidência do STJ, j. 5 abr. 2016.
41. BRASIL. Superior Tribunal de Justiça. REsp n. 1.642.997/RJ, Minª. Relª. Nancy Andrighi, 3ª Turma, j. 12 set. 2017.
42. BRASIL. Superior Tribunal de Justiça. REsp n. 1.698.647/SP, Minª. Relª. Nancy Andrighi, 3ª Turma, j. 6 fev. 2018.

43. BRASIL. Supremo Tribunal Federal, Tribunal Pleno, ADPF 130, Rel. Min. Carlos Ayres Brito, j. 30 abr. 2009.
44. BRASIL. Supremo Tribunal Federal. *Tema 987*. Disponível em: <https://portal.stf.jus.br/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=5160549&numeroProcesso=1037396&classeProcesso=RE&numeroTema=987>. Acesso em 5 dez. 2024.
45. BRASIL. Supremo Tribunal Federal. Tema 533. Disponível em: <https://portal.stf.jus.br/jurisprudenciarepercussao/verAndamentoProcesso.asp?incidente=5217273&numeroProcesso=1057258&classeProcesso=RE&numeroTema=533>. Acesso em 5 dez. 2024.
46. BRASIL. Supremo Tribunal Federal. RE n. 1.037.396/SP, Min. Rel. Dias Toffoli, Plenário, voto do Relator em 5 dez. 2024.
47. BRASIL. Supremo Tribunal Federal. RE n. 1.057.258/MG, Min. Rel. Luiz Fux, Plenário, voto do Relator em 11 dez. 2024.
48. BRASIL. Supremo Tribunal Federal. RE n. 201819/RJ. Minª. Relª. Ellen Gracie, Rel. p/ Acórdão Min. Gilmar Mendes, 2ª Turma, j. 11 out. 2005.
49. BRASIL. Supremo Tribunal Federal. RE n. 639138/RS. Min. Gilmar Mendes, Rel. p/ Acórdão Min. Edson Fachin, Plenário, j. 18 ago. 2020 (Tema 452/STF).
50. BRASIL. Tribunal Superior Eleitoral. *Resolução 23/2022*. Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022. Acesso em 9 dez. 2024.
51. BREGA, Gabriel Ribeiro. *A regulação de conteúdo nas redes sociais: uma breve análise comparativa entre o NetzDG e a solução brasileira*. Revista GV, São Paulo, v.19, e2305, 2023, p. 14. Disponível em: <https://doi.org/10.1590/2317-6172202305>. Acesso em 22 dez. 2024).
52. BUDISH, Ryan; WOOLERY, Liz; e BANKSTON, Kevin. *The Transparency Reporting Toolkit: Survey & Best Practice Memos for Reporting on U.S. Government Requests for User Information*. New America's Open Technology Institute (OTI) and Harvard University's Berkman Center for Internet & Society. Disponível em: <https://www.newamerica.org/oti/policy-papers/the-transparency-reporting-toolkit/>. Acesso em 20 jan. 2025.
53. BUTANTAN. *Queda nas taxas de vacinação no Brasil ameaça a saúde das crianças*. Butantan, publicado em 7 mar. 2022. Disponível em: <https://butantan.gov.br/noticias/queda-nas-taxas-de-vacinacao-no-brasil-ameaca-a-saude-das-criancas>. Acesso em 29 nov. 2024.
54. CABRAL, Antônio do Passo. Repensando a autotutela: conceito e limites no direito brasileiro. *Revista de Processo*, abril, 2024, vol. 350, pp. 21-47, ano 49.
55. CARLUCCI, Manoela. Por telefone, Lula e Macron conversam sobre decisão da Meta. CNN Brasil, publicado em 12 jan. 2025. Disponível em: <https://www.cnnbrasil.com.br/politica/por-telefone-lula-e-macron-conversam-sobre-decisao-da-meta/>. Acesso em 13 jan. 2025.
56. CARNEIRO, Paulo Cezar Pinheiro. *Acesso à justiça: juizados especiais cíveis e ação civil pública — uma nova sistematização da teoria geral do processo*. Rio de Janeiro: Forense, 2000.
57. CARVALHO, Luísa. *Ponto a ponto: entenda o voto de Barroso sobre o artigo 19 do Marco Civil da Internet*. JOTA, publicado em 20 dez. 2024. Disponível em: <https://www.jota.info/stf/do-supremo/ponto-a-ponto-

entenda-o-voto-de-barroso-sobre-o-artigo-19-do-marco-civil-da-internet>. Acesso em 7 mar. 2025.

58. CELESTE, Edoardo. Digital Constitutionalism: Mapping the Constitutional Response to Digital Technology's Challenges. *HIIG Discussion Paper Series*, n. 02, 2018, p. 15. Disponível em: <https://ssrn.com/abstract=3219905>. Acesso em 5 jan. 2025.
59. CGI.br. Consulta sobre Regulação de Plataformas Digitais. Disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 18 fev. 2025.
60. COALIZÃO DE DIREITOS NAS REDES. *Relatório de Referências Internacionais em regulação de plataformas digitais: bons exemplos e lições para o caso brasileiro, publicado em 23 abr. 2024*. Disponível em: <https://direitosnarede.org.br/2024/04/23/coalizao-direitos-na-rede-lanca-o-relatorio-referencias internacionais-em-regulacao-de-plataformas-digitais-bons-exemplos-e-licoes-para-o-caso brasileiro/>. Acesso em 10 fev. 2025.
61. COALIZÃO DIREITOS NAS REDES. *Órgão independente de supervisão das plataformas é essencial, mas não pode ser Anatel*. Publicado em 28 abr. 2023. Disponível em: <https://direitosnarede.org.br/2023/04/28/orgao-independente-de-supervisao-das-plataformas-e-essencial-mas-nao-pode-ser-anatel/?ref=nucleo.jor.br>. Acesso em 16 fev. 2025.
62. COALIZÃO DIREITOS NAS REDES. *PL 2630: propostas da CDR para uma lei efetiva e democrática*. Disponível em: <https://direitosnarede.org.br/2020/09/01/pl-2630-propostas-da-cdr-para-uma-lei-efetiva-e-democratica/>. . Acesso em 10 fev. 2025.
63. COALIZÃO DIREITOS NAS REDES. PL 2630: *Regulação pública democrática das plataformas é fundamental, com instituições autônomas e participativas.* Publicado em 28 abr. 2023. Disponível em: <https://direitosnarede.org.br/2023/04/28/pl-2630-regulacao-publica-democratica-das-plataformas-e-fundamental-com-instituicoes-autonomas-e-participativas/>. Acesso em 14 fev. 2025.
64. COMISSÃO ESPECIAL DE DIREITO DIGITAL DO CONSELHO FEDERAL DA OAB. *Ofício nº 001/2023 – CEDD/OAB*. Publicado em 13 mai. 2023. Disponível em: <https://nucleo.jor.br/content/files/2023/05/Oficio-Sistema-brasileiro-de-regulac-a-o-de-plataformas-digitais.pdf>. Acesso em 16 fev. 2025.
65. COMISSÃO EUROPEIA. *The Digital Services Act*. Disponível em: <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en>. Disponível em 15 fev. 2025.
66. CROKER, Andrew; GEBHART, Gennie; MACKEY, Aaron; OPSAHL, Kurt; TSUKAYAMA, Hayley; WILLIAMS, Jamie Lee; YORK, Jillian C. *Who Has Your Back? Censorship Edition 2019*. Disponível em: <https://www.eff.org/files/2019/06/11/whyb_2019_report.pdf>. Acesso em 20 jan. 2025.
67. C-SAN.GOV. *Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*. Publicado em 10 abr. 2018. Disponível em: <https://www.c-span.org/program/senate-committee/facebook-ceo-mark-zuckerberg-hearing-on-data-privacy-and-protection/500690>. Acesso em 27 dez. 2024.

68. CUETO, José Carlos. *Rússia restringe acesso ao Facebook e Twitter após ataques à Ucrânia*. BBC News Brasil, 26 fev. 2022. Disponível em: <https://www.bbc.com/portuguese/articles/cq5zydp59j7o>. Acesso em 13 jan. 2025.
69. CURZI, Yasmin. ZINGALES, Nicolo. GASPAR, Walter. LEITÃO, Clara. COUTO, Natália. REBELO, Leandro. OLIVEIRA, Maria Eduarda. *Nota técnica do Centro de Tecnologia e Sociedade da FGV Direito Rio sobre o substitutivo ao PL 2630/2020*. Rio de Janeiro: FGV Direito Rio, 2021.
70. DE LIMA, Venício A. *Da cultura do silêncio ao direito à comunicação*. Observatório da Impresa, publicado em 22 nov. 2011. Disponível: <https://www.observatoriodaimprensa.com.br/feitos-desfeitas/da-cultura-do-silencio-ao-direito-a-comunicacao/>. Acesso em 26 jan. 2025.
71. DOUEK, Evelyn. Content Moderation as Systems Thinking. *Harvard Law Review*, vol. 136, 2022, p. 43-44. Disponível em: <https://ssrn.com/abstract=4005326>. Acesso em 17 fev. 2025.
72. DOUEK, Evelyn. Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review*, v. 121, n. 3, 2021, p. 759-834, p. 791. Disponível em: <https://ssrn.com/abstract=3679607>. Acesso em 9 dez. 2024.
73. ELETRIC FRONTIER FOUNDATION et al. *Princípios de Manila sobre Responsabilidade de Provedores*. Disponível em: <https://manilaprinciples.org/index.html>. Acesso em 15 dez. 2024.
74. ELETRIC FRONTIER FOUNDATION. *Padrões de Direitos Humanos como Linhas de Base para a Regulação e Prestação de Contas das Plataformas: uma contribuição para o debate brasileiro*. Disponível em: <https://www.eff.org/files/2023/07/07/padroes_de_direitos_humanos_como_linhas_de_base_para_a_regulacao_e_prestacao_de_contas_das_plataformas_pt-br.pdf>. Acesso em 18 fev. 2025.
75. ESTADOS UNIDOS DA AMÉRICA. *Communications Decency Act*. Disponível em: <https://en.wikisource.org/wiki/Telecommunications_Act_of_1996#TITLE_V%E2%80%94OBSCENITY_AND_VIOLENCE>. Acesso em 22 dez. 2024.
76. ESTADOS UNIDOS DA AMÉRICA. *Digital Millennium Copyright Act*. Disponível em: <https://www.congress.gov/bill/105th-congress/house-bill/2281/text>. Acesso em 05 dez. 2024.
77. FACEBOOK. *Padrões de Comunidade*. Disponível em: <https://transparency.meta.com/pt-br/policies/community-standards/child-sexual-exploitation-abuse-nudity/>. Acesso em 26 dez. 2024.
78. FALCÃO, Márcio. Moraes dá cinco dias para o X explicar lives de contas bloqueadas pela Justiça. *TV Globo*. Rio de Janeiro, 22 abr. 2024. Disponível em: <https://g1.globo.com/politica/noticia/2024/04/22/moraes-da-cinco-dias-para-o-x-explicar-lives-de-contas-bloqueadas-pela-justica.ghtml>. Acesso em 17 jan. 2024.
79. FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*. São Paulo: Perspectiva, 2020.
80. FARINHO, Domingos Soares. Delimitação do espectro regulatório de redes sociais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 43-104

81. FISS, Owen M. The irony of free speech. Cambridge: Harvard University Press, 1998,
82. FORBES BRASIL. Weibo, o Twitter chinês, já é mais valioso que o original. *Forbes*, publicado em 12 out. 2016. Disponível em: <https://forbes.com.br/negocios/2016/10/weibo-o-twitter-chines-ja-e-mais-valioso-que-o-original/>. Acesso em 13 jan. 2025.
83. FORBES. Brasil é o terceiro país com mais usuários do YouTube em 2023. *Forbes*, publicado em 10 mai. 2023. Disponível em: <https://forbes.com.br/forbes-tech/2023/05/brasil-e-o-terceiro-pais-com-mais-usuarios-do-youtube-em-2023/>. Acesso em 9 dez. 2024.
84. FREIRE, Paulo. *Pedagogia do oprimido*, 17ª. Ed. Rio de Janeiro, Paz e Terra, 1987. Disponível: <https://pibid.unespar.edu.br/noticias/paulo-freire-1970-pedagogia-do-oprimido.pdf/view>. Acesso em 26 jan. 2025.
85. G1. Cambridge Analytica se declara culpada em caso de uso de dados do Facebook. *G1*, publicado em 9 de jan. de 2019. Disponível: <https://g1.globo.com/economia/tecnologia/noticia/2019/01/09/cambridge-analytica-se-declara-culpada-por-uso-de-dados-do-facebook.ghtml>. Acesso em 27 dez. 2024.
86. G1. Instagram restrito na Rússia: entenda a importância da rede social para o país de Putin. *G1*, publicado 12 mar. 2022. Disponível em: <https://g1.globo.com/tecnologia/noticia/2022/03/12/instagram-restrito-na-russia-entenda-a-importancia-da-rede-social-para-o-pais-de-putin.ghtml>. Acesso em 13 jan. 2025.
87. GLOBAL PARTNERS DIGITAL. *Content Regulation in the Digital Age*. OHCHR, sem data. Disponível em: <GlobalPartnersDigital.pdf>. Acesso em 14 fev. 2025.
88. GLOBAL PARTNERS DIGITAL. *Content regulation laws threaten our freedom of expression. We need a new approach*. Publicado em 15 mai. 2018. Disponível em: <https://www.gp-digital.org/content-regulation-laws-threaten-our-freedom-of-expression-we-need-a-new-approach/>. Acesso em 14 fev. 2025.
89. GOMES, Alessandra; ANTONIALLI, Dennys; OLIVA, Thiago. *Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet?*. Disponível em: <https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/>. Acesso em 28 dez. 2024
90. GRAGNANI, Juliana; SENRA, Ricardo. Movimento antivacina é criminoso, diz Drauzio Varella. *BBC News*, 26 jun. 2019. Disponível em: <https://www.bbc.com/portuguese/geral-48780905>. Acesso em 9 dez. 2024.
91. GUIDO, Gabriela. 8 de janeiro e atentado a bomba legitimam STF para julgar regulamentação das redes sociais. *Valor*, publicado em 27 nov. 2024. Disponível em: <https://valor.globo.com/politica/noticia/2024/11/27/8-de-janeiro-e-atentado-a-bomba-legitimam-stf-para-julgar-regulamentacao-das-redes-sociais.ghtml>. Acesso em 28 fev. 2025.
92. HESSE, Konrad. *Elementos de Direito Constitucional da Alemanha Federal*. Luís Afonso Heck (tradutor). Porto Alegre: S.A. Fabris, 1998.
93. HOLANDA, Marianna. Lula convoca reunião sobre Meta e diz que um cidadão não pode ferir soberania da nação. *Folha de S. Paulo*, publicado em

12 jan. 2025. Disponível em: <https://www1.folha.uol.com.br/poder/2025/01/lula-convoca-reuniao-sobre-meta-e-diz-que-um-cidadao-nao-pode-ferir-soberania-da-nacao.shtml>. Acesso em 13 jan. 2025.

94. INSTAGRAM. *Termos de utilização*. Publicado em 26 jul. 2022. Disponível em: <https://help.instagram.com/581066165581870/?locale=pt_PT&hl=pt>. Acesso em 13 jan. 2025.
95. ITS RIO. *7 reflexões para o futuro do debate sobre moderação de conteúdo em plataformas digitais*. Publicado em 19 mar. 2025. Disponível em: https://itsrio.org/pt/publicacoes/moderacao-conteudo-plataformas-digitais-its-rio/. Acesso em 7 abr. 2025.
96. ITS RIO. *9 pontos de atenção sobre a PL das Fake News (PL 2630/2020)*. Publicado em 31 mar. 2022. Disponível em: <https://itsrio.org/wp-content/uploads/2022/04/9-pontos-de-aten%C3%A7%C3%A3o-sobre-o-PL-das-Fake-News-PL-2630_20.pdf>. Acesso em 14 fev. 2025.
97. ITS RIO. *10 pontos de atenção sobre a PL das Fake News (PL 2630/2020)*, publicado em 28 mar. 2022. Disponível em: <https://itsrio.org/wp-content/uploads/2022/03/10-pontos-de-atencao-sobre-o-PL-das-Fake-News-PL-2630_20.pdf>. Acesso em 13 fev. 2025.
98. ITS RIO. *Regulação de Plataformas Digitais*. Disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 16 fev. 2025.
99. ITS RIO. *Regulação de Plataformas Digital*. Audiência pública da CGI.br. Disponível em: <https://dialogos.cgi.br/documentos/debate/consulta-plataformas/>. Acesso em 16 fev. 2025.
100. KELLER, Daphne. *Platform Transparency and the First Amendment* (March 3, 2023), p. 12. Disponível em: <https://ssrn.com/abstract=4377578>. Acesso em 3 jan. 2025
101. KELLER, Daphne; LEERSSEN, *Paddy. Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation* (December 16, 2019). Forthcoming, N. Persily & J. Tucker, Social Media and Democracy: The State of the Field and Prospects for Reform (Cambridge University Press), p. 36. Disponível em: <https://ssrn.com/abstract=3504930>. Acesso em 29 nov. 2024.
102. KHAN, Irene. *Disinformation and freedom of opinion and expression. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. United Nations General Assembly, Human Rights Council, Forty-seventh session, A/HRC/47/25. Publicado em 13 abr. 2021. Disponível em: <https://digitallibrary.un.org/record/3925306>. Acesso em 13 fev. 2025.
103. KLONICK, Kate. The New Governors: The People, Rules, and Processes Governing Online Speech (March 20, 2017). 131 *Harvard Law Review*, pp. 1598-1599. Disponível em: <https://ssrn.com/abstract=2937985>. Acesso em 24 nov. 2024.
104. KONDER, Carlos Nelson de Paula; SOUZA, Amanda Guimarães Cordeiro de. Onerosidade do acesso às redes sociais. *Revista de Direito do Consumidor*, ano 28, vol. 121, jan.-fev./2019, pp. 185-212:
105. LESSIG, Lawrence. *Code: version 2.0*. Basic Books, 2006.REAUTERS. Instagram é bloqueado na China em meio a protestos em Hong Kong. G1,

29 set. 2014. Disponível em: <https://g1.globo.com/tecnologia/noticia/2014/09/instagram-e-bloqueado-na-china-em-meio-protestos-em-hong-kong.html>. Acesso em 13 jan. 2025.

106. LYNGAAS, Sean. Eleição dos EUA 'vê quantidade sem igual de desinformação', diz chefe de segurança cibernética. *CNN*, 4 nov. 2024. Disponível em: <https://www.cnnbrasil.com.br/internacional/eleicoes-nos-eua-2024/eleicao-dos-eua-ve-quantidade-sem-igual-de-desinformacao-diz-chefe-de-seguranca-cibernetica/>. Acesso em 29 nov. 2024.
107. MACCARTHY, *Mark. Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry* (February 12, 2020). Transatlantic Working Group, 2020, p. 7. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615726>. Acesso em 11 jan. 2025
108. MARANHÃO, Juliano; CAMPOS, Ricardo. Fake News e autorregulação regulada das redes sociais no Brasil: fundamentos constitucionais. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021, pp. 341-355.
109. MARTIN, Ana. Vacina contra Covid no PNI: 'Fato ou Fake' desmentiu dezenas de informações falsas; veja 10 delas. *Portal G1*, publicado em 19 jul. 2023. Disponível em: <https://g1.globo.com/sp/sao-carlos-regiao/noticia/2023/07/19/vacina-contra-covid-no-pni-fato-ou-fake-desmentiu-dezenas-de-informacoes-falsas-veja-10-delas.ghtml>. Acesso em 9 dez. 2024.
110. MENDES, Gilmar Ferreira; BRANCO, Paulo Gustavo Gonet. *Curso de Direito Constitucional*. 9ª ed. São Paulo: Saraiva, 2014.
111. MENDES, Gilmar Ferreira; FERNANDES, Victor Oliveira. Constitucionalismo digital e jurisdição constitucional: uma agenda de pesquisa para o caso brasileiro. *Revista Brasileira de Direito*, Passo Fundo, vol. 16, n. 1, pp. 1-33, jan.-abr., 2020.
112. META. *Cross-lingual pretraining sets new state of the art for natural language understanding*. Publicado em 4 fev. 2019. Disponível em: <https://ai.meta.com/blog/cross-lingual-pretraining/>. Acesso em 28 dez. 2024.
113. MONTEIRO, Artur Pericles Lima; CRUZ et al. Francisco Brito; SILVEIRA, Juliana Fonteles da; e VALENTA, Mariana G. *Armadilhas e caminhos na regulação da moderação de conteúdos: Diagnósticos & Recomendações #5*. São Paulo: InternetLab, 2021.
114. MORAES, Alexandre de. *Direito Constitucional*. 40ª Edição, 2024. Rio de Janeiro: Atlas, 2024 (*ebook*).
115. MULHOLLAND, Caitlin. Responsabilidade civil indireta dos provedores de serviço de internet e sua regulação no Marco Civil da Internet. In José Renato Gaziero Cella, Aires Jose Rover, Valéria Ribas Do Nascimento (Coords.). *Direito e novas tecnologias*. Florianópolis: CONPEDI, 2015, pp. 479-502.
116. MULHOLLAND, Isabella. Quem anunciará onde fake news e conteúdo tóxico correm soltos?. *Meio&Mensagem*, publicado em 14 dez. 2023. Disponível em: <https://www.meioemensagem.com.br/opiniao/quem-anunciara-onde-fake-news-e-conteudo-toxico-correm-soltos>. Acesso em 13 jan. 2025.

117. NITRINI, Rodrigo Vidal. *Liberdade de Expressão nas Redes Sociais: o problema jurídico da remoção de conteúdo pelas plataformas*. Belo Horizonte: Dialética, 2021.
118. NOELLE-NEUMANN, Elisabeth. *A espiral do silêncio – Opinião pública: nosso tecido social*. Cristian Derosa (Trad.). Florianópolis: Estudos Nacionais, 2017.
119. NOGUEIRA, Gustavo Santana; NOGUEIRA, Suzane de Almeida Pimentel. O sistema de múltiplas portas e o acesso à justiça no brasil: perspectivas a partir do novo código de processo civil. *Revista de Processo*, vol. 276, pp. 505-522, fev. 2018.
120. ORWELL, George. *1984*. Traduzido por Karla Lima. São Paulo: Principis, 2021.
121. OVERSIGHT BOARD. *Discurso do General Brasileiro*. Publicado em 22 jun. 2023. Disponível em: Brazilian general's speech | Oversight Board. Acesso em 16 fev. 2025.
122. PASQUETTO, Irene V, et. al. Tackling misinformation: what researchers could do with social media data. *Harvard Kennedy School (HKS) Misinformation Review*, 2020. Disponível em: <https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/>. Acesso em 21 jan. 2025.
123. PEREIRA DE LIMA, Cíntia Rosa; FRANCO DE MORAES, Emanuele Pezati; PEROLI, Kelvin. O necessário diálogo entre o Marco Civil da Internet e a Lei Geral de Proteção de Dados para a coerência do sistema de responsabilidade civil diante das Novas Tecnologias. In: *Responsabilidade civil e novas tecnologias*; coordenado por Guilherme Magalhães Martins. Nelson Rosenvald – Indaiatuba, SP. Editora Foco, 2020.
124. PORTAL G1. Twitter, Facebook e Instagram bloqueiam contas de Trump temporariamente. Portal G1, publicado em 6 jan. 2025. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/2021/01/06/twitter-diz-que-conta-de-trump-ficara-bloqueada-por-12-horas.ghtml>. Acesso em 9 fev. 2025.
125. RÁDIO USP. Alta lucratividade é o que mantém o mercado digital de fake news. *Rádio USP*, publicado em 5 dez. 2022. Disponível em: <https://jornal.usp.br/radio-usp/alta-lucratividade-e-o-que-mantem-o-mercado-digital-de-fake-news/>. Acesso em 23 jan. 2025.
126. PSAFE. *Relatório da segurança digital no Brasil: terceiro trimestre – 2018*. Disponível em <https://www.psafe.com/dfndr-lab/pt-br/relatorio-da-seguranca-digital>. Acesso em 30 nov. 2024.
127. REUTERS. Musk diz que plataforma X atingiu novo recorde de usuários mensais. *Reuters*, publicado em 29 jul. 2023. Disponível em: <https://www.infomoney.com.br/negocios/musk-diz-que-plataforma-x-atingiu-novo-recorde-de-usuarios-mensais/>. Acesso em 9 dez. 2024.
128. RODRIGUES JUNIOR, Otávio Luiz. Artigo 5º, incisos IV ao IX. In: Paulo Bonavides e Jorge Miranda e Walber de Moura Agra (Coordenadores Científicos). Francisco Bilac Pinto Filho e Otavio Luiz Rodrigues Junior (Coordenadores Editoriais). *Comentários à Constituição Federal de 1988*. Rio de Janeiro: Forense, 2009.
129. SAAD, Beth; MALAR, João Pedro. Mudanças da Meta ilustram a nova e perigosa era das redes sociais. *Folha de S. Paulo*, publicado em 12 jan. 2025.

Disponível em: <https://www1.folha.uol.com.br/opiniao/2025/01/mudancas-da-meta-ilustram-a-nova-e-perigosa-era-das-redes-sociais.shtml>. Acesso em 13 jan. 2025.

130. SALLES, Raquel Bellini de Oliveira. *Autotutela nas relações contratuais*. Rio de Janeiro: Editora Processo, 2019.
131. SARLET, Ingo Wolfgang; MARINONI, Luiz Guilherme; MITIDIERO, Daniel. *Curso de direito constitucional*. 4ª ed. São Paulo: Saraiva, 2015 (ebook).
132. SARMENTO, Daniel. Comentários ao art. 5.º, incisos IV, V e IX. In: CANOTILHO, J. J. Gomes; MENDES, Gilmar Ferreira; SARLET, Ingo Wolfgang, STRECK, Lenio Luiz (coord.). *Comentários à Constituição do Brasil*. São Paulo: Saraiva/Almedina, 2013 (ebook).
133. SARTOR, Giovanni; LOREGGIA, Andrea. *The impact of algorithms for online content filtering or moderation: Upload filters*. Luxembourg: European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, 2020, p. 43. Disponível em: <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf>. Acesso em 26 dez. 2024.
134. SCHREIBER, Anderson. Marco Civil da Internet: Avanço ou retrocesso? A responsabilidade civil por dano derivado do conteúdo gerado por terceiro. In: LUCCA, Newton de; SIMÃO FILHO, Adalberto; LIMA, Cíntia Rosa Pereira. *Direito e Internet III: Marco Civil da Internet, Lei nº 12.965/2014*. t. II. São Paulo: Quartier Latin, 2015, pp. 277-305.
135. SCHROEPFER, Mike. Community standards report. *Meta*, publicado em 13 de nov. de 2019. Disponível em: <https://ai.meta.com/blog/community-standards-report/>. Acesso em 28 dez. 2024.
136. SILVA, José Afonso da. *Curso de direito constitucional positivo*. 22ª ed. São Paulo: Malheiros, 2003.
137. SILVA, Orlando. *Relatório final do Relator Deputado Orlando Silva*. Apresentado em 31 de março de 2022, no Plenário da Câmara dos Deputados. Disponível em: <https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=PRLP+1+%3D%3E+PL+2630/2020>. Acesso em 15 dez. 2024.
138. SILVA, Rodrigo da Guia. Notas sobre o cabimento do direito de retenção: desafios da autotutela no direito privado. Civilistica.com. Rio de Janeiro, a. 6, n. 2, 2017. Disponível em: <http://civilistica.com/notas-sobre-o-cabimento-do-direito-de-retencao/>. Acesso em 12 jan. 2024.
139. SILVA, Virgílio Afonso da. Direitos fundamentais e relações entre particulares. Revista Direito GV, v.1, n. 1, p. 173-180, maio, 2005, p. 174. Disponível em: <https://periodicos.fgv.br/revdireitogv/article/view/35274/34067>. Acesso em 5 jan. 2025.
140. SILVEIRA, Janaína. WeChat: o app faz tudo que mudou a vida dos chineses. *Veja*, 23 nov. 2018. Disponível em: <https://veja.abril.com.br/mundo/wechat-o-app-faz-tudo-que-mudou-a-vida-dos-chineses>. Acesso em 13 jan. 2025.
141. SPRING, Mariana. Como usuários do X ganham milhares de dólares espalhando fake news sobre eleição dos EUA. *BBC News Brasil*, 30 out.

2024. Disponível em: <https://www.bbc.com/portuguese/articles/c937q4p7g09o>. Acesso em 17 jan. 2025.

142. SUNDFELD, Carlos Ari; ROSILHO, André. A governança não estatal da internet e o direito brasileiro. *RDA*, Rio de Janeiro, v. 270, pp. 41/79, set./dez. 2015. Disponível em: <https://periodicos.fgv.br/rda/article/view/58737/57530>. Acesso em 18 fev. 2025.
143. SUPREMA CORTE DOS ESTADOS UNIDOS DA AMERICA. *Moody, Attorney General of Florida, et al. v. NetChoice, LLC, DBA NetChoice, et al.* Disponível em: <https://www.supremecourt.gov>. Acesso em 13 jan. 2025.
144. SUPREMA CORTE DOS ESTADOS UNIDOS DA AMERICA. *Murthy, Surgeon General, et al. v. Missouri et al*. 06 jun. 2024. Disponível em: <www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf> . Acesso em 13 jan. 2025.
145. SUPREMA CORTE DOS ESTADOS UNIDOS. *Stratton Oakmont, Inc. v. Prodigy Services Co.* 1995 WL 323710 at *5 (N.Y. Supr. Ct. May 23, 1995.
146. SUPREMA CORTE DOS ESTADOS UNIDOS DA AMERICA. *Zeran v. America Online, Inc.*, 985 F. Supp. 1124 (E.D. Va. 1997).
147. SUPREMA CORTE DOS ESTADOS UNIDOS. *Cubby, Inc. v. CompuServe Inc.*, 776 F. Supp. 135. United States District Court, S.D. New York. Oct. 29, 1991.
148. TERRA, Aline de Miranda Valverde Terra. Inafastabilidade da jurisdição e autotutela: o exemplo da cláusula resolutiva expressa. *Revista Eletrônica de Direito Processual*, Rio de Janeiro, ano 13, vol. 20, n. 3, pp. 1-19, set./ dez. 2019.
149. THEODORO JÚNIOR, Humberto; ANDRADE, Érico. Novas perspectivas para atuação da tutela executiva no direito brasileiro: autotutela executiva e "desjudicialização" da execução. *Revista de Processo*, vol. 315, pp. 109-158, mai. 2021.
150. TOFFOLI, Dias. Fake News: desinformação e liberdade de expressão. In: Georges Abboud, Nelson Nery Jr. e Ricardo Campos (Org.). *Fake News e Regulação*. 3ª ed. São Paulo: Revista dos Tribunais, 2021.
151. TÔRRES, Fernanda Carolina. O direito fundamental à liberdade de expressão e sua extensão. *Revisão de Informação Legislativa*. Senado Federal. Ano 50 Número 200 out./dez. 2013, p. 62. Disponível em: <https://www12.senado.leg.br/ril/edicoes/50/200/ril_v50_n200_p61.pdf/>. Acesso em 27 dez. 2024.
152. TOTVS. *Hash: o que é, importância e como funciona*. Publicado em 28 out. 2024. Disponível em: https://www.totvs.com/blog/gestao-para-assinatura-de-documentos/hash-assinatura-digital/ . Acesso em 26 dez. 2024.
153. UNESCO. *Safeguarding freedom of expression and access to information: guidelines for a multistakeholder approach in the context of regulating digital platforms*. Publicado em 27 abr. 2023. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en>. Acesso em 13 fev. 2025.
154. UNIÃO EUROPEIA. *2022 Strengthened Code of Practice on Disinformation*. Disponível em: <https://digital-

strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>. Acesso em 15 fev. 2025.

155. UNICEF. *3 em cada 10 crianças no Brasil não receberam vacinas que salvam vidas, alerta UNICEF. 27 abr. 2022*. Disponível em: <https://www.unicef.org/brazil/comunicados-de-imprensa/3-em-cada-10-criancas-no-brasil-nao-receberam-vacinas-que-salvam-vidas>. Acesso em 29 nov. 2024.
156. UOL. *Zuckerberg diz que Meta vai acabar com checagem de fatos, cita censura e manda recado ao STF*. Disponível em: <https://www.youtube.com/watch?v=nJQt3DLQqQ0>. Acesso em 13 jan. 2025
157. VALENTE, Mariana Giogetti. A liberdade de expressão na internet: Da utopia à era das plataformas. In: FARIA, José Eduardo Faria. *Liberdade de expressão e as novas mídias*. São Paulo: Perspectiva, 2020.
158. VENTURI, Thaís Goveia Pascoaloto. *A construção da responsabilidade civil preventiva no direito civil contemporâneo*. (Doutorado). Faculdade de Direito da Universidade Federal do Paraná, Paraná, 2012.
159. VIEIRA RAMOS, Carlos Eduardo. *O Direito das Plataformas: Procedimento, legitimidade e constitucionalização na regulação privada da liberdade de expressão na internet*. (Mestrado) Faculdade de Direito da Universidade de São Paulo, São Paulo, 2020.
160. WARDLE, Claire; DERAKHSHAN, Hossein (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe report, DGI (2017)09. Disponível em: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>. Acesso em 29 nov. 2024.