PRISMA JURÍDICO
e-ISSN: 1983-9286
ISSN: 1677-4760
COPE Member since 2020 JM14464

Recebido em: 09 ago. 2021 / **Aprovado em:** 17 maio 2022
Editor: Prof. Dr. José Fernando Vidal de Souza
Editora Assistente: Heloisa Corrêa Meneses
Processo de Avaliação: *Double Blind Review*
http://doi.org/10.5585/prismaj.v21n1.20573
Artigo

Check for updates

A moderação de conteúdo em massa por plataformas privadas de redes sociais

Large-scale content moderation by private social media platforms

Álerton Emanuel Poletto
Faculdade Meridional - IMED - RS
Mestre em Direito
alertonep@gmail.com

Fausto Santos de Morais
Faculdade Meridional - IMED - RS
Doutor e Mestre em Direito
faustosmorais@gmail.com

Resumo: As redes sociais exercem um controle em massa de todo conteúdo que é publicado pelos usuários, por atividade de algoritmos. A fim de compreender se a atividade moderadora está de acordo com o ordenamento jurídico, precede abordar o fundamento jurídico para a atividade. Ainda, verifica-se que os mecanismos tradicionais de proteção dos cidadãos não fazem frente aos novos desafios do ambiente digital. Esses mecanismos são verificados no marco regulatório das redes sociais e nas propostas legislativas de regulamentação da moderação de conteúdo. A partir da análise das tentativas regulatórias da moderação de conteúdo, apresenta-se uma ideia de regulação, a fim de assegurar a manifestação dos usuários e a diminuição de remoção de conteúdo, prejudicial ao espaço público. A metodologia em relação à linha de abordagem refere-se ao método fenomenológico-hermenêutico, pela compreensão da moderação de conteúdo institucionalizada no ordenamento jurídico. Do ponto de vista procedimental, é do tipo exploratória e monográfica.

Palavras-chave: moderação de conteúdo; redes sociais; regras de uso; inteligência artificial; marcos legislativos.

Abstract: Social media exert a massive control over all content that is posted by users, by algorithmic activity. In order to understand if the moderating activity is in accordance with the legal system, it is necessary to address the legal basis for the activity. Still, it appears that the traditional mechanisms for protecting citizens do not face the new challenges of the digital environment. These mechanisms are verified in the regulatory framework of social networks and in the legislative proposals for regulating content moderation. From the analysis of regulatory attempts to moderate content, an idea of regulation is presented, in order to ensure the expression of users and the reduction of removal of content, which is harmful to the public space. The methodology in relation to the line of approach refers to the phenomenological-hermeneutic method, by understanding the institutionalized content moderation in the legal system. From a procedural point of view, it is exploratory and monographic.

Keywords: content moderation; social media; terms of use; artificial intelligence; legislative frameworks.

Para citar este artigo
ABNT NBR 6023:2018

POLETTO, Álerton Emmanuel; MORAIS, Fausto Santos de. A moderação de conteúdo em massa por plataformas privadas de redes sociais. *Prisma Jurídico,* São Paulo, v. 21, n. 1, p. 108-126, jan./jun. 2022. http://doi.org/10.5585/prismaj.v21n1.20573

Introdução

O presente estudo delimita-se dentro do tema da afetação da atividade de moderação do conteúdo sobre os usuários, dentro de redes sociais de plataformas privadas. Assim, tem-se como objeto de pesquisa uma análise do controle de conteúdo publicado nas redes sociais pelos usuários, a fim de compreender se as ações do Facebook, Instagram e Twitter, como bloqueio, remoção e filtro de conteúdo publicado em sua plataforma social, estão de acordo com a ordem jurídica brasileira.

Firma-se que o Estado tradicional não possui condições de promover a proteção dos usuários-cidadãos no ambiente digital, uma vez que as plataformas de redes sociais não encontram barreiras jurídicas ou limites à sua atividade, ficando a sua própria autorregulação.

Contudo, frente às tendências internacionais de regular as plataformas digitais, o Brasil discute propostas legislativas para ordenar esses entes privados e garantir direitos mínimos aos usuários, urgindo a necessidade da discussão acadêmica desses marcos legislativos da regulação da internet e das redes sociais digitais.

O objetivo central do consiste em apresentar como a moderação de conteúdo é operacionalizada pelas plataformas privadas, identificando suas implicações, à luz do ordenamento jurídico.

A verificação desse fenômeno nas plataformas sociais exige do Direito uma resposta eficiente. Posto isso, a pesquisa se baseia no problema de como o ordenamento jurídico brasileiro regula ou busca regular a atividade de moderação de conteúdo pelas redes sociais em contraponto com a experiência internacional, se essa regulação efetivamente garante a proteção dos usuários e como se pode aperfeiçoar.

Metodologicamente, cumpre delinear como ocorre a prática da moderação de conteúdo pelas redes sociais. Posterior a isso, é imprescindível verificar o fundamento jurídico que legitima o Facebook, Instagram e Twitter em controlar suas plataformas, para, por fim, explorar as regulações internacionais sobre a moderação e as propostas e tentativas no Brasil.

Do ponto de vista procedimental, é do tipo exploratória e monográfica, partindo da exploração de obras de vários autores, os quais debatem a temática abordada, configurando-se o corpus da pesquisa realizada. No que tange à linha de abordagem e elaboração da temática refere-se ao método fenomenológico-hermenêutico, pela compreensão da moderação de conteúdo institucionalizada no ordenamento jurídico.

A moderação representa a atividade de controle do conteúdo que circula nas redes sociais, como Facebook e Twitter, pelas próprias plataformas. Tendo em vista que diariamente

o fluxo de publicações de circulam nas redes sociais ultrapassa a casa do bilhão, as plataformas implementaram sistemas inteligentes para operar essa atividade. Assim, a moderação é uma atividade em grande escala de remoção e restrição de conteúdo operacionalizada por algoritmos, devendo ao Direito analisar as implicações jurídicas.

Desse modo, o presente inicia-se pelo delineamento da atividade de moderação de conteúdo pelas redes sociais, com a consequente explicitação do fundamento jurídico que legitima essa atividade das plataformas e, por fim, uma análise das regulações das redes sociais pela ordem constitucional brasileira.

1 A prática de moderação de conteúdo pelas redes sociais

Para compreender a prática de moderação de conteúdo pelas redes sociais é preciso abordar o que venha a ser a moderação, as técnicas de moderação, as suas características e os problemas decorrentes desse processo. Essas questões serão apresentadas nesta seção.

Inicialmente, cumpre explanar o que seja a moderação de conteúdo e delinear suas características. Uma plataforma totalmente "aberta" sem qualquer regulação ou moderação de conteúdo é uma utopia da ideia democrática, uma vez que todas as plataformas moderam conteúdos e impõem regras aos usuários, até porque, seria insustentável a utilização do serviço (GILLESPIE, 2018, p. 5).

Nessa esteira, tem-se como moderação de conteúdo a atividade das plataformas digitais de triagem de todo o conteúdo publicado pelos usuários, bem como marcação desse conteúdo como o que pode ou não estar no ambiente digital, de acordo com as regras da empresa, com o intuito de proteger os usuários e prevenir abusos.

A moderação é um serviço indispensável para as plataformas, no que tange à proteção dos usuários, ocultação de conteúdo ilegal, bem como para a organização de todo o conteúdo que circula na rede. Nesta senda, a moderação torna-se um processo necessário no controle de conteúdo pornográfico, obsceno, violento, ilegal, abusivo e de ódio. Dessa forma, resta analisar seis técnicas de moderação que as plataformas exercem na organização do fluxo de conteúdo dos autores para os leitores.

A primeira forma é a exclusão do conteúdo (*deletion*) que opera removendo a publicação que esteja em desacordo com as regras impostas pela plataforma (GRIMMELMANN, 2015, p. 56). O exemplo mais polêmico representa a exclusão pelo Facebook da histórica foto da Guerra do Vietnã, na qual uma criança corre desnuda. A postagem foi marcada pela Inteligência Artificial moderadora como inapropriada em razão da nudez e excluída da página,

desconsiderando o contexto histórico da imagem. Após diversas críticas, o Facebook, ao revisar a decisão da moderação, percebeu o erro e republicou a postagem.

Na sequência, tem-se a moderação de edição (*editing*), de forma autoexplicativa, é a alteração de conteúdo que varia em corrigir erros de digitação alterando a essência de um post, isto é, o moderador rejeita a verdade do autor e as substituiu pela dele próprio (GRIMMELMANN, 2015, p. 59).

Outra operação de moderação é anotação (*annotation*), a qual se traduz pela adição de informações ou por comentários no conteúdo (GRIMMELMANN, 2015, p. 59). Toma-se como exemplo o sistema de feedback dos compradores e vendedores nas plataformas de e-commerce na qual usuários escrevem comentários e críticas, bem como os Likes e Comentários no Facebook e Instagram.

Por seguinte, a síntese (*synthesis*) é a combinação de transformação de conteúdo. O Wikipédia é o principal exemplo dessa organização de fluxo, uma vez que os usuários podem promover pequenas e heterogêneas mudanças sintetizadas em uma enciclopédia digital (GRIMMELMANN, 2015, p. 59).

A filtragem de conteúdo (*filtering*) é uma ferramenta não destrutiva como a exclusão, mas que pode resultar no mesmo efeito prático. A operacionalização desse fluxo sustenta-se como um motor de busca que filtra o conteúdo e apresenta apenas os mais relevantes aos leitores (GRIMMELMANN, 2015, p. 59). A exemplificação mais precisa acontece na linha do tempo das redes sociais do Facebook e Instagram que apresentam no início da página o conteúdo mais relevante e o menos relevante no final, resultando em uma "exclusão" deste último.

Por último, a formatação (*formatting*) é o modelo de moderação de conteúdo que apresenta uma tipografia melhorada ao exibir informações aos leitores, tornando a legibilidade ordenada e rápida (GRIMMELMANN, 2015, p. 59).

Em suma, a moderação de conteúdo representa a atividade das redes sociais de filtrar o conteúdo publicado pelos usuários, com a marcação desse conteúdo se pode ou não estar no ambiente digital e a consequente remoção, ocultação e restrição do conteúdo problemático. Ressalta-se que a operacionalização dessa atividade pode ser procedida pelo esforço humano ou artificial, sendo uma tendência pela moderação através de sistemas inteligentes devido ao volume de fluxo informacional.

A partir desses fatores, a moderação do fluxo de conteúdo auxilia os leitores a ver apenas o conteúdo que possuem preferência e ocultar os menos interessantes ou ilegais, sob o pálio de melhorar a experiência de uso da rede. Todavia, essas ferramentas podem se tornar mecanismos

de manipulação nas mãos de moderadores não engajados com as diretrizes das plataformas e a ética ou pela operação de inteligências artificiais comprometidas.

Importa ressaltar que ao passo que as plataformas de comunicação crescem o caos e a discórdia crescem paralelamente. Isso acontece em razão de que os usuários possuem a necessidade de se expressarem, seja de forma inspiradora ou repreensiva, desde que "eu possa dizer onde outros vão me ouvir" (GILLESPIE, 2018, p. 5).

A moderação de conteúdo não conseguirá prevenir todos os casos das patologias da vida online, por isso é necessário estabelecer limites toleráveis, sem, até mesmo, elevar os custos jurídicos-sociais da moderação de conteúdo a um nível inaceitável (GRIMMELMANN, 2015, p. 53).

Em escala global, diariamente bilhões de conteúdos são publicados na plataforma incorrendo na tarefa do moderador em decidir o que fica ou não online. Frente a isso, sistemas de Inteligência Artificial (IA) são implementados para automatizar a triagem para identificar e remover conteúdo prejudicial, como spam e nudez.

Os sistemas de IA tendem a remover muito do conteúdo publicado na rede considerado spam e nudez, todavia, para outras categorias, como discurso de ódio e assédio, não respondem positivamente, uma vez que ainda requerem discernimento humano de contexto e nuance para a compreensão (BARRET, 2020, p. 3). Portanto, requer-se uma atividade conjunta entre a IA e humanos, principalmente está na função revisora para diminuir os erros da moderação de determinados conteúdos.

Apenas um adendo, "o emprego de ferramentas de inteligência artificial que interagem por meio de mídias sociais exige que se reflita sobre os requisitos éticos que devem acompanhar o desenvolvimento desse tipo de tecnologia" (MAGRANI, 2019, p. 220). Considerando que o que é publicado nas redes sociais exerce influência tanto nas relações dentro quanto fora do ambiente digital, cumpre realizar o alerta acerca das questões éticas do implemento de ferramentas automatizadas, como enviesamento e opacidade dos algoritmos.

Outro ponto a ser alertado decorre de que sistemas com inteligência artificial utilizados na moderação de conteúdo tendem a desenvolver uma postura proativa de remoção de conteúdo, necessitando, portanto, da transparência dos resultados da IA quanto à taxa de acerto da moderação de conteúdo contrário às regras das plataformas de aplicação.

A moderação de conteúdo é realizada pelas plataformas com base na detecção de marcação de conteúdo que estaria violando uma regra da rede social. Nesse ponto convém destacar que as plataformas disponibilizam de um sistema de revisão e marcação de conteúdo potencialmente lesivo ou abusivo para a utilização dos próprios usuários da rede social.

Essa marcação é realizada pelo esforço tecnológico com o auxílio humano na revisão. A tecnologia do Facebook detecta e remove proativamente a grande maioria dos conteúdos violadores antes que os usuários os denunciem. Assim, as plataformas de redes sociais utilizam de ferramentas que marcam o conteúdo como violador, classificam conforme a diretriz violada e, sendo caso, removem do ambiente digital (FACEBOOK, 2021c, online).

Não obstante, problemas legais, como arbitrariedade da decisão de moderação e exclusão de conteúdo, implicam questionamentos jurídicos quanto à possibilidade de haver contraditório ao usuário, à inferência de alguma justificação ou fundamentação da decisão da moderação e, ainda, à recorribilidade da decisão.

Em uma tentativa de promover níveis mínimos de governança nas plataformas de redes sociais foi editado uma carta de princípios para servir como ponto de partida no diálogo com empresas envolvidas com moderação de conteúdo. Os Princípios de Santa Clara (*The Santa Clara Principles*) sobre transparência e responsabilidade na moderação de conteúdo no âmbito digital, buscam delinear níveis mínimos de transparência significativa e responsabilidade pelo respeito dos direitos dos usuários, com o intuito de mitigar o impacto dos conteúdos nocivos propagados na internet (THE SANTA CLARA PRINCIPLES, 2020).

O primeiro dos três princípios propostos, denominado *Numbers*, corresponde a publicação dos números de postagens removidas ou contas suspensas por violação às diretrizes de conteúdos da plataforma. O segundo princípio, *Notice*, traduz o dever de notificação do usuário sobre o motivo pelo qual sua conta foi suspensa ou conteúdo removido. E, por fim, o *Appeal*, estabelece que as empresas devem oportunizar mecanismos de recursos aos usuários acerca das decisões de remoção de conteúdo ou suspensão de contas (THE SANTA CLARA PRINCIPLES, 2020).

A iniciativa e adesão à carta descrita representa um esforço das próprias plataformas no comprometimento com os direitos dos usuários. Todavia, somente esses princípios não garantem a proteção dos direitos e do ambiente digital, requer-se um esforço conjunto entre as plataformas, sociedades civis e Estados para promover níveis mínimos de proteção, através uma regulação das redes sociais.

Nessa esteira, na perspectiva jurídica, as plataformas digitais acenderam como um espaço no qual direitos são exercidos, incidindo no aumento da moderação do conteúdo nas redes sociais pelo aumento do fluxo informacional. Posto isso, cumpre explanar o fundamento jurídico para essa atividade.

2 O fundamento jurídico para a moderação de conteúdo

Na presente seção será abordado o fundamento jurídico que legitima a moderação de conteúdo pelo Facebook, Instagram e Twitter, bem como evidenciado as regras da plataforma para determinar o que pode ou não ser publicado. A delimitação nessas plataformas se deu pela popularidade das redes, números de usuários e, principalmente, casos controversos com decisões de moderação.

Inicialmente, insta abordar o Facebook por ser a rede social com maior número de usuários. A relação entre o usuário e o Facebook é selada pelo aceite adesivo dos Termos de Serviço ao realizar o cadastro na plataforma. Os Termos importam em diretrizes que regem o uso do Facebook, juntamente com outros termos específicos para determinados serviços fornecidos pela plataforma, como Política de Cookies e Política de Dados.

Reconhecendo a importância do Facebook como um local de comunicação, com o intuito de eliminar dos serviços todo tipo de abuso, a empresa desenvolveu um conjunto de Padrões da Comunidade que detalham o que é ou não permitido no Facebook, baseado nas áreas de tecnologia, segurança pública e direitos humanos (FACEBOOK, 2021, online).

Malgrado a plataforma reconhecer que a internet cria novas e maiores oportunidades de abusos, o Facebook assegura o comprometimento com a livre expressão, sendo as limitações baseadas em valores de autenticidade, segurança, privacidade e dignidade (FACEBOOK, 2021, online).

Por mais que a plataforma objetiva a promoção da comunicação online, as limitações de conteúdo possuem, muito mais, um interesse comercial. Esse interesse "é uma razão central para que empresas privadas de mídia social proíbam manifestação que não é ilegal e, portanto, limitem a liberdade de expressão muito além do exigido pela lei" (HARTMANN; SARLET, 2019, p. 96).

Vale explanar, além disso, que os Padrões da Comunidade classificam os conteúdos publicados em 24 tópicos dentro de cinco diretrizes, dentre elas: I. Comportamento violento e criminoso; II. Segurança; III. Conteúdo questionável; IV. Integridade e autenticidade; e, V. Com respeito à propriedade intelectual (FACEBOOK, 2021, online).

A primeira diretriz, deste modo, engloba todo o conteúdo de violência e incitação, organizações e indivíduos perigosos, coordenação de danos e divulgação de crime, divulgação de produtos controlados, bem como fraude e dolo. No que tange à diretriz da Segurança, os conteúdos são subdivididos em: automutilação e suicídio; exploração sexual, abuso ou nudez

infantil; exploração sexual de adultos; bullying e assédio; exploração humana; e violações de privacidade e direitos de privacidade de imagem (FACEBOOK, 2021, online).

Já na diretriz dos conteúdos questionáveis se encontram as publicações marcadas por discurso de ódio, violência e conteúdo explícito, nudez adulta e atividades sexuais, abordagem sexual, como também conteúdo cruel e insensível. Outra importante diretriz diz respeito à integridade e autenticidade, a qual compreende a integridade da conta e identidade autêntica, spam, segurança cibernética, comportamento não autêntico, notícias falsas, mídia manipulada e perfil memorial. Por fim, em atenção aos direitos autorais, marcas comerciais e outros direitos legais alheios, a última diretriz concerne aos problemas relativos à propriedade intelectual que possam decorrer do uso da rede social (FACEBOOK, 2021, online).

Por fazer parte do grupo econômico do Facebook Inc., o Instagram compartilha diretrizes de políticas de serviços com o Facebook, mas de forma independente. Dentre as Diretrizes da Comunidade do Instagram, destaca-se: direitos de propriedade intelectual; proibição de nudez; limitar a disseminação de spam; proibição de conteúdo violento e criminoso; limitação de conteúdo que contenha ameaça real, discurso de ódio ou ataques individuais (INSTAGRAM, 2021, online).

Especialmente quanto às contas de figuras públicas na plataforma, o Facebook possui regras específicas para tratamento contra violações, dentre as quais são considerados o engajamento da conta, alcance e influência da exposição da personalidade pública, tendo em vista o temor de representarem um risco maior de danos ao violar as políticas.

A definição de quem seriam as figuras públicas no Facebook traduz-se em "funcionários do governo estadual e nacional, candidatos políticos para esses cargos, pessoas com mais de um milhão de fãs ou seguidores nas redes sociais e pessoas que recebem cobertura noticiosa substancial" (Tradução livre. FACEBOOK, 2021d, online).

Outrossim, destaca-se uma nova medida de restrição de contas de figuras públicas durante distúrbios civis, quando o perfil público atuar na rede social de modo a incitar ou promover eventos violentos ou agitação civil. De tal modo, a duração da restrição imposta pelo Facebook pode variar de um mês a 2 anos a critérios da empresa, conforme o grau de violação (FACEBOOK, 2021d, online).

Nessa senda, o Facebook e Instagram são serviços de redes sociais diferentes, mas compartilham da mesma ideologia da política de padrões da comunidade, ressalvada as peculiaridades de cada rede social.

No Twitter, as regras de moderação de conteúdo são diferentes. Por ser uma plataforma de rede social que permite aos usuários compartilharem publicamente mensagens curtas, há

uma maior liberdade quanto aos conteúdos, sendo vedados somente os que estejam em desacordo com as Regras do Twitter (TWITTER, 2021, online).

Dentre as regras de conteúdo do Twitter, no que tange à diretriz da segurança, são proibidos os conteúdos: violentos; terroristas; de exploração sexual de menores; abuso ou assédio; discurso de ódio; suicídio e automutilação; mídia sensível; e finalidade ilegal. Em relação à privacidade, é proibido publicar informações pessoais privadas, bem como nudez não consensual. Por sua vez, as regras de autenticidade impõem proibição contra: spam e manipulação da plataforma, seja para propagar mais informações ou interferência na integridade cívica; identidade falsa; compartilhamento de mídias sintéticas e manipuladas; direitos autorais (TWITTER, 2021, online).

Não obstante, o Twitter possui um sistema de distinção e marcação com etiquetas em contas de mídia governamental e afiliada ao Estado, da qual consta a informação "sobre o país ao qual a conta está afiliada e se é utilizada por um representante do governo ou por uma entidade de mídia afiliada ao Estado" (TWITTER, 2021, online).

É mister evidenciar que as redes sociais possuem três funções públicas chaves para o bom funcionamento do ambiente digital. Primeiramente, as redes sociais facilitam a participação pública na arte, política e cultura; em segundo lugar, a mídia social organiza conversas públicas para que as pessoas possam se encontrar e se comunicar facilmente; por fim, as mídias sociais moderam a opinião pública (BALKIN, 2020, p. 6)

Com efeito, havendo a verificação da violação dos Padrões da Comunidade, podem decorrer consequências que variam conforme a gravidade e com o histórico do usuário na plataforma, das quais "podemos notificar alguém por uma primeira violação, mas se a pessoa persistir na violação de nossas políticas, podemos restringir sua possibilidade de publicar no Facebook ou mesmo desativar seu perfil" (FACEBOOK, 2021, online).

Frente a uma violação da política de conteúdo do Twitter, como primeiro ato, a plataforma irá notificar o usuário para que remova o conteúdo, podendo, ainda, temporariamente impedir o acesso à conta. Caso a violação aconteça novamente, a conta fica permanentemente suspensa. Sobreleva notar que a plataforma dispõe de mecanismos de marcação de mídia sensível ao publicar, a fim de minimizar o contato dos usuários com esse tipo de conteúdo, além da função de denunciar a conta ou tweet (TWITTER, 2019, online).

No que tange ao Instagram, a rede social vem desenvolvendo diversas ferramentas de segurança, privacidade e denunciação de abusos. Assim, cumpre referenciar a indicação de conteúdo sensível, bem como a possibilidade de denunciação, bloqueio e controle de publicação, conta, comentário ou mensagem (INSTAGRAM, 2021a, online).

Essas medidas podem ser verificadas no exemplo do presidente americano Donald Trump e do presidente Jair Bolsonaro. No decorrer do ano de 2020 ambos presidentes tiveram publicações removidas das principais redes sociais com a justificativa de que causavam desinformação acerca da COVID-19, podendo gerar prejuízos sanitários ao combate da pandemia (SENRA, 2020, online).

Destacado o histórico do presidente norte-americano em violar as diretrizes em diversas oportunidades, a última decisão de bloquear a conta de Trump por tempo indeterminado foi a mais dura punição imposta pelo Facebook, relacionada com a moderação dos perfis do presidente americano, devido ao ataque no capitólio por parte de seus apoiadores extremistas (G1, 2021, online).

Em suma, as regras de utilização de cada rede social importam em diretrizes para orientar a moderação de conteúdo do que estaria violando ou não as regras de uso da plataforma, devendo, portanto, essas regras estarem de acordo com o ordenamento jurídico do Estado o qual o usuário é jurisdicionado.

3 Regulação da moderação de conteúdo

Da moderação de conteúdo decorrem diversas implicações negativas aos usuários e seus direitos exercidos no ambiente digital. Posto isso, cumpre apresentar, nesta seção, os pontos nocivos da moderação de conteúdo, problematizando quanto à legitimidade das plataformas em julgar os conteúdos.

Ainda, a fim de cumprir com objetivo de verificar o marco legal da moderação de conteúdo no Brasil, é relevante expor as propostas de regulação da moderação de conteúdo na experiência internacional, para, posteriormente, contrapor com as propostas legislativas brasileiras e as discussões jurisprudenciais acerca da temática.

A partir da atividade de moderação de conteúdo decorrem implicações como limitação da liberdade de expressão, exposição a conteúdos sensíveis, discricionariedade do moderador, controle do discurso público, remoção de conteúdo em massa. Desta feita, perfaz necessário questionar como se afere a legitimidade do conteúdo, de modo a indagar se cabe à plataforma julgar o conteúdo e até que ponto é legítimo a plataforma filtrar o conteúdo e julgar o conteúdo, bem como ter como padrão a remoção do conteúdo.

Em síntese, as regras de conteúdo permitido nas redes sociais importam em diretrizes para orientar a moderação de conteúdo do que estaria violando ou não as regras de uso da plataforma, contudo, os termos propostos pelas plataformas são padrões mínimos para a

moderação de conteúdo ilegal, necessitando de uma regulação nacional, de acordo com ordenamento jurídico.

Apenas fazendo um adendo, é importante ressaltar que os termos de serviço do Facebook seguem as práticas de moderação de conteúdo dos Estados Unidos da América, uma vez ser o país sede da plataforma de aplicação online.

Nessa esteira, em uma primeira linha de orientação, resta imperioso realizar uma investigação dos modelos e regulamentações das redes sociais no âmbito internacional, a fim de analisar as respostas jurídicas para a proteção de direito fundamentais relacionado a publicação online de conteúdo ilegal e moderação pelas plataformas online.

Metodologicamente, cumpre delimitar a investigação do quadro regulatório e estruturas de políticas relacionado com conteúdo ilegal online na Europa e nos Estados Unidos da América, uma vez que essas nações são matrizes orientativas para as discussões jurídicas no Brasil.

No ordenamento norte-americano, as discussões jurídicas acerca dos conflitos online recaem sobre a seção 230, do *United States Code*, denominada de Proteção para bloqueio privado e triagem de material ofensivo. Em suma, a seção preceitua uma proteção dos sites de ações judiciais caso um usuário postar algo ilegal, salvo exceções contra violações de direitos autorais, conteúdo relacionado a trabalho sexual e violações da lei criminal (ESTADOS UNIDOS DA AMÉRICA, 2021, online).

Malgrado a Primeira Emenda da constituição americana proíba a qualquer restrição estatal nos discursos, alargando a liberdade de expressão, a emenda protege e legitima as regras de moderação de conteúdo do Facebook que restringem publicações que contenham discurso de ódio, mesmo que seja legalmente permitido nos Estados Unidos.

Na experiência europeia, a Alemanha, a França, o Reino Unido e Portugal editaram respectivas leis e políticas nacionais sobre moderação de conteúdo ilegal e prejudicial online. A Lei alemã de Fiscalização da Rede (NetzDG) foi decretada em junho de 2017 para melhorar a fiscalização das disposições penais existentes na Internet e, mais especificamente, nas redes sociais, cominando penas e multas às empresas de tecnologia (ALEMANHA, 2017).

O Parlamento francês, em maio de 2020, aprovou a lei de combate ao discurso de ódio na Internet, a chamada Lei de Avia, cujo escopo é regulamentar a moderação de conteúdo ilegal para as empresas de plataformas online. Não obstante, o Conselho Constitucional Francês, tribunal constitucional da França, declarou a inconstitucionalidade de algumas disposições da Lei de Combate ao Ódio Online, tidas como contrárias à tendencia do constitucionalismo

digital, no que tange à exigência de uma ordem administrativa da polícia, bem como o requisito do conteúdo ser manifestamente ilegal (ARTICLE 19, 2020, online).

Por sua vez, o Reino Unido lançou o *White Paper on Online Harms* de 2019 regulamentando os mecanismos para retirar o conteúdo ilegal online, bem como introduzindo a noção de um dever de cuidado, por meio de uma moderação proativa e constante das postagens por parte das plataformas (REINO UNIDO, 2019, online).

Em relação à regulação portuguesa, em 2021 foi editada a Carta Portuguesa de Direitos Humanos na Era Digital, a fim de tutelar direitos, liberdades e garantias no ciberespaço. Especificamente em relação à moderação de conteúdo, a carta dispões de uma seção que afirma o direito à liberdade de expressão nas suas diversas formas, bem como preceitua que medidas que visam impedir o acesso ou remover conteúdo deverão ser objetos de lei especial (PORTUGAL, 2021, online).

O conjunto de normas proposto pela Comissão Europeia, denominado de Lei dos Serviços Digitais (*Digital Services Act*), dispõe acerca de regras sobre as obrigações e a responsabilidade das plataformas na prestação de serviços digitais além-fronteiras, a fim de assegurar um elevado nível de proteção a todos os usuários-cidadãos da União Europeia. A proposta segue em discussão no Parlamento Europeu (COMISSÃO EUROPEIA, 2021).

Na perspectiva legislativa brasileira, há a observância do Marco Civil da Internet (Lei 12.965/2014), o qual estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. No que tange às redes sociais, a aplicação da legislação importa na responsabilidade por danos decorrentes de publicações dos usuários (BRASIL, 2014, online).

Entretanto, em sede de recurso extraordinário, a aplicabilidade da normativa encontra-se em discussão pelo Tema de Repercussão Geral n° 533, no Supremo Tribunal Federal, para a fixação da tese sobre o 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 (SUPREMO TRIBUNAL FEDERAL, 2021, online).

Já a Lei Geral de Proteção de Dados (LGPD) dispõe sobre o tratamento de dados pessoais, com o intuito de proteger os direitos fundamentais de liberdade, privacidade e livre desenvolvimento da personalidade nos ambientes digitais (BRASIL, 2018, online). Em consonância com a recente entrada em vigor da LGPD, o Facebook teve que atualizar os termos de coletas de dados e solicitar o consentimento dos usuários para atender aos dispositivos legais, no que tange ao tratamento de dados e privacidade (FACEBOOK, 2021a, online).

Um aspecto relevante da lei diz respeito às decisões automatizadas. A LGPD dispõe que as decisões tomadas unicamente por meios automatizados são passíveis de recurso pelo usuário,

o qual deverá ser informado que a decisão foi procedida por sistema automatizado, bem como quais os critérios e procedimentos foram observados na decisão (BRASIL, 2018, online).

Acerca da temática, encontra-se em discussão o Projeto de Lei nº 21/2020, o qual propõe estabelecer princípios, direitos e deveres para o uso de inteligência artificial no país, devendo ser observado para a atividade de moderação de conteúdo proativa por sistemas de IA. O Projeto de Lei intenta criar um marco legal para o desenvolvimento e uso da IA tanto pelo poder público, quanto por empresas, entidades diversas e pessoas físicas (BRASIL, 2020a, online).

De forma técnica, o texto legislativo prevê e impõem a observância do respeito aos direitos humanos, aos valores democráticos, a igualdade, a não discriminação, a pluralidade, a privacidade dos dados, bem como a transparência sobre o uso e operacionalidade da IA. Uma medida inovadora apresentada no Projeto de Lei nº 21/2020 estabelece a formação de relatório de impacto da Inteligência Artificial, com a descrição da tecnologia utilizada, a fim de manter adequação com padrões institucionalizados no PL (BRASIL, 2020a, online).

Ainda, em discussão legislativa, há o Projeto de Lei n° 2630/20, instituído de Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet, a qual propõe normas, diretrizes e mecanismos de transparência para provedores de redes sociais e de serviços de mensageria privada, com a finalidade de garantir segurança e ampla liberdade de expressão, comunicação e manifestação do pensamento (BRASIL, 2020, online).

O projeto dispõe de uma seção específica para regulamentar o procedimento da moderação de conteúdo dos provedores de redes sociais que ofertam serviços ao público brasileiro ou possuam estabelecimento no país. Ademais, o PL 2630 cria um órgão responsável pelo acompanhamento das medidas institucionalizadas no texto legal, denominado de Conselho de Transparência e Responsabilidade na Internet, além de permitir a criação de instituição de autorregulação regulada pelas próprias plataformas (BRASIL, 2020, online).

No texto original do projeto de lei, a moderação de conteúdo encontra uma regulamentação jurídica no artigo 12. O dispositivo impõe aos provedores de redes sociais o dever de acesso à informação e à liberdade de expressão dos usuários em relação aos termos de uso de cada plataforma, bem como a disponibilização de ferramentas de recurso das decisões de moderação e estabelecimento de um procedimento (BRASIL, 2020, online).

Outrossim, o artigo determina a notificação do usuário, quando ocorrer uma denúncia ou uma medida seja aplicada em razão dos termos de uso sobre conteúdo ou contas, acerca da fundamentação, do processo de análise e da aplicação da medida, bem como sobre o procedimento de contestação da decisão (BRASIL, 2020, online).

A legislação traz uma exceção à prévia notificação do usuário frente à verificação de: dano imediato de difícil reparação; risco à segurança de informações ou do usuário; violação a direitos de crianças e adolescentes; crimes resultantes de preconceito de raça ou de cor; ou grave comprometimento da funcionalidade da plataforma (BRASIL, 2020, online).

Potencialmente, das produções legislativas observam-se duas correntes quanto à vinculação da obrigação das plataformas de aplicação na moderação de conteúdo. Na obrigação de resultado, sob influência da responsabilização objetiva, a legislação impõe uma responsabilidade pelo resultado, isto é, há um incentivo grande à remoção de conteúdos, uma vez que caso um conteúdo potencialmente lesivo cause um dano a um usuário a plataforma seria responsabilidade. Outrossim, além de aumentar a incidência de remoção de conteúdo, as legislações que firmam a obrigação de resultado acabam por ampliar o poder das redes sociais e a própria censura privada (HARTMANN, 2021, online).

Em contraposto, a obrigação de meio busca estabelecer critérios e obrigações na remoção de conteúdos para as plataformas de aplicação, principalmente no que tange à transparência da moderação, seja dos dados ou do próprio procedimento, bem como assegurar direitos e garantias aos usuários contra abusos da própria plataforma (HARTMANN, 2021, online).

Nesta senda, a falta de limites que caracteriza o ambiente digital e as dificuldades técnicas de implementação implicam controle mais rígido baseado em um modelo de sancionamento. Para Sarlet e Sales, somente por meio do empoderamento dos usuários para o uso responsável da internet e a disseminação de uma cultura de respeito e tolerância, bem como, "a elaboração de um Tratado Internacional para a Proteção dos Direitos Humanos na Internet e um código de ética comum para a internet, podem vir a ter resultados mais eficazes e duradouros" (Tradução livre. SARLET; SALES, 2021, online).

Outrossim, a eficácia dos marcos normativos da internet condiciona-se a sua aceitação e compromisso dos Estados e das empresas de aplicação, através de uma "rede de cooperação e política multinível e a interação jurídica, em um contexto de autorregulação regulada das plataformas de mídia social" (Tradução livre. SARLET; SALES, 2021, online).

Portanto, insta referenciar a atuação do Estado em buscar promover a proteção eficiente dos usuários-cidadãos nas redes sociais, frente aos novos paradigmas das relações sociais provenientes do ambiente digital. As discussões legislativas no Brasil caminham com a tendência internacional de uma regulação autorregulada, bem como a instituição de agências reguladoras para exercer o controle e fiscalização das plataformas.

Contudo tem-se que as legislações postas e as discussões legislativas, nos termos propostos, não garantem uma participação livre dos usuários contra arbítrios das próprias plataformas. As decisões de moderação não sofrem alterações com as regulações, permitindo a prática de remoção proativa em massa de conteúdo, não resolvendo os problemas que implicam aos usuários, decorrentes dessa atividade.

Considerações finais

O objetivo da presente pesquisa consistiu na identificação das implicações jurídicas à atividade de moderação de conteúdo pelas plataformas privadas, de modo que ficou evidenciado o problema proposto com base nas regulamentações das redes sociais e da moderação de conteúdo.

O controle de conteúdo pelas plataformas é imprescindível para o funcionamento das redes sociais. A moderação encontra seu fundamento nas diretrizes e padrões determinados pelas redes sociais, os quais, de certo modo, observam os preceitos básicos dos direitos exercidos no ambiente digital, como a privacidade e a liberdade de expressão.

Todavia, a atividade de moderação é complexa e por ser regulada através de diretrizes modestas de aplicabilidade global, encontram diversas questões jurídicas quando em conflito com abuso aos direitos dos usuários, uma vez que as plataformas de redes sociais não encontram barreiras jurídicas ou limites à sua atividade, ficando à sua própria autorregulação.

Por outro lado, tem-se que o Estado não possui condições de promover a proteção dos usuários-cidadãos no ambiente digital pelos métodos tradicionais, urgindo uma necessidade de adequação digital. Portanto, é um dever do Estado em impor e garantir a proteção dos usuários-cidadãos, estabelecendo uma regulação e o controle da atividade das plataformas.

Assim, com vistas ao projeto de lei exposto, busca-se uma tentativa de estabelecer princípios e normas positivadas em busca da proteção do ambiente digital. As propostas e discussões legislativas precisam institucionalizar uma obrigação legal de meio que garanta direitos mínimos aos usuários contra arbítrios das redes e também imponha limites ao poder das plataformas de aplicação. Tal discussão encontra materialização parcial no projeto de Lei 2630/2020 que busca instituir uma agência regulatória responsável pelo acompanhamento das medidas institucionalizadas, além de consentir pela criação de instituição de autorregulação regulada pelas próprias plataformas.

Todavia, acredita-se que as regulações nos termos propostos não asseguram a livre participação dos usuários em um ambiente de debate público contra arbítrios das plataformas.

Uma legislação ideal busca promover mecanismos que inibam a discricionariedade das decisões de moderação de conteúdo, impedindo uma censura privada, por meio da não remoção de conteúdo.

Malgrado a remoção de alguns conteúdos sensíveis ser justificável (tais como nudez adulta e infantil, spam, violador de direitos autorais e violência explícita), firma-se o caminho para as obrigações de meio entre o usuário e a plataforma na qual o conteúdo não é retirado do ambiente digital, mas marcado como potencialmente lesivo ao usuário.

Isto é, a legislação deveria assegurar a permanência da maioria dos conteúdos ao invés de incentivar a remoção. Uma ferramenta que poderia ser implementada é a de marcação do conteúdo como sensível, inverídico, potencial abusivo. Ressalta-se que as plataformas já disponibilizam desse sistema de marcação de conteúdo, todavia, escolhem em remover em grande escala.

Assim, essas medidas importariam na redução da remoção de conteúdos e na fixação de avisos nas publicações. Por fim, essas medidas tornam o ambiente de mínima intervenção no exercício de direitos dos usuários e de segurança dos usuários pela própria plataforma. Possibilitando, portanto, o direito do usuário de estar presente na rede social e decidir qual conteúdo vai consumir.

Referências

ALEMANHA. Ministério Federal da Justiça e Defesa do Consumidor. **Network Enforcement Act** (Netzwerkdurchsetzungsgesetz -NetzDG), de 01 de setembro de 2017. Disponível em: https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html. Acesso em: 02 mar. 2021.

BALKIN, Jack M., How to Regulate (and Not Regulate) social media. **Knight Institute Occasional Paper Series**, n. 1 mar. 2020. Disponível em: https://knightcolumbia.org/content/how-to-regulate-and-not-regulate-social-media. Acesso em: 13 nov. 2020.

BARRETT, Paul M. **Who Moderates the Social Media Giants?** A Call to End Outsourcing. Nova Iorque: New York University, 2020.

BRASIL. Câmara dos Deputados. Projeto de Lei n° 2.630, de 04 de fevereiro de 2020. **Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet**. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735&fichaAmigavel=nao. Acesso em: 03 mar. 2021.

BRASIL. Câmara dos Deputados. Projeto de Lei n° 21, de 03 de julho de 2020a. **Estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá outras providências**. Disponível em:

https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=1853928. Acesso em: 03 jul. 2021.

BRASIL. Lei 13.709 de 14 de agosto de 2018. **Lei Geral de Proteção de Dados (LGPD)**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 03 mar. 2021.

BRASIL. Lei nº 12.965, de 23 de abril de 2014. **Marco Civil da Internet**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 01 mar. 2021.

BRASIL. Supremo Tribunal Federal. **Tema de Repercussão Geral n° 533.** 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. Relator: Min. Luiz Fux. Disponível em: http://www.stf.jus.br/portal/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=4155926&numeroProcesso=660861&classeProcesso=ARE&numeroTema=533. Acesso em: 02 mar. 2021.

ESTADOS Unidos da América. **United States Code**. Disponível em: https://uscode.house.gov/browse/prelim@title47/chapter5/subchapter2/part1&edition=prelim. Acesso em: 03 mar. 2021.

FACEBOOK e Instagram bloqueiam conta de Trump por tempo indeterminado. **G1**, [s.l.], 07 de jan. 2021. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2021/01/07/facebook-bane-conta-de-donald-trump-por-tempo-indeterminado-diz-mark-zuckerberg.ghtml. Acesso em: 01 mar. 2021.

FACEBOOK**. Padrões da Comunidade**. [s.l.] 2021. Disponível em: https://www.facebook.com/communitystandards/. Acesso em: 01 mar. 2021.

FACEBOOK. **Política de dados**. [s.l.] 2021a. Disponível em: https://www.facebook.com/policy.php. Acesso em: 01 mar. 2021.

FACEBOOK. **Termos de Serviço**. [s.l.] 2020. Disponível em: https://www.facebook.com/legal/terms/update. Acesso em: 01 mar. 2021.

FACEBOOK. **Transparency Center**. [s.l.] 2021c. Disponível em: https://transparency.fb.com/ Acesso em: 16 jun. 2021.

FRANCE: Constitutional Council declares French hate speech 'Avia' law unconstitutional. **ARTICLE 19**,18 jun. 2020. Disponível em: https://www.article19.org/resources/france-constitutional-council-declares-french-hate-speech-avialaw-unconstitutional/. Acesso em: 03 mar. 2021.

GILLESPIE, Tarleton. **Custodians of the Internet**: plataforms, contente moderation, and yhe hidder decisions that shape social media. [s.l.] Yale University Press, 2018.

GRIMMELMANN, James. The Virtues of Moderation. **Yale Journal of Law & Technology**. v. 17, p. 41-109, 2015.

HARTMANN PEIXOTO, Fabiano. **Direito e Inteligência Artificial**: Referenciais básicos com comentários à resolução CNJ 332/2020. Brasília: DR.IA, 2020.

HARTMANN, Ivar Alberto M. Fake News, Pandemia e Moderação de Conteúdo. Entrevista concedida a Fausto Santos de Morais e Lucas Carini. **IAJUS TALK**. [s.l.] 2021. Disponível em: https://www.youtube.com/watch?v=ijeas6USNK8&t=774s

HARTMANN, Ivar Alberto M.; SARLET, Ingo Wolfgang. DIREITOS FUNDAMENTAIS E DIREITO PRIVADO: A PROTEÇÃO DA LIBERDADE DE EXPRESSÃO NAS MÍDIAS SOCIAIS. **Direito Público**, [S.l.], v. 16, n. 90, dez. 2019. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/3755 Acesso em: 07 mar. 2021.

INSTAGRAM. **Diretrizes da Comunidade**. [s.l.] 2021. Disponível em: https://help.instagram.com/477434105621119/?helpref=hc_fnav&bc[0]=Ajuda%20do%20Instagram&bc[1]=Central%20de%20Privacidade%20e%20Seguran%C3%A7a Acesso em: 18 mar. 2021.

INSTAGRAM. **Keeping Instagram a safe and supportive place.** [s.l.] 2021a. Disponível em: https://about.instagram.com/community/safety Acesso em: 18 mar. 2021.

MAGRANI, Eduardo. **Entre dados e robôs**: ética e privacidade na era da hiperconectividade. 2 ed. Porto Alegre: Arquipélago Editorial, 2019.

PORTUGAL. Lei n. 27/2021 de 17 de maio de 2021. **Carta Portuguesa de Direitos Humanos na Era Digital**. Disponível em: https://dre.pt/application/conteudo/163442504 Acesso em: 24 jun. de 2021.

REINO UNIDO. Government Digital Service. **Online Harms White Paper** de abril de 2019. Disponível em: https://www.gov.uk/government/consultations/online-harms-white-paper. Acesso em: 02 mar. 2021.

SARLET, Ingo Wolfgang; SALES, Gabrielle Bezerra. FREEDOM OF EXPRESSION AND HATE SPEECH REGULATION IN SOCIAL MEDIA PLATFORMS: CONSIDERATIONS ON THE EXAMPLE OF THE SO-CALLED “GERMAN NETWORK ENFORCEMENT ACT” ("NETZWERKDURCHSETZUNGSGESETZ"). **REVISTA DE DERECHO CONSTITUCIONAL EUROPEO**. Granada, n. 35, jan-jun de 2021. Disponível em: http://www.ugr.es/~redce/REDCE35/articulos/12_SARLET_SALES.htm Acesso em: 01 ago. 2021.

SENRA, Ricardo. Após Twitter, Facebook e Instagram excluem vídeo de Bolsonaro por 'causar danos reais às pessoas'**. BBC News**, 2020. Disponível em: https://www.bbc.com/portuguese/brasil-52101240 Acesso em: 20 mar. 2021.

THE SANTA CLARA PRINCIPLES on Transparency and Accountability in Content Moderation. Disponível em: https://santaclaraprinciples.org/. Acesso em: 01 mar. 2021.

TWITTER. **As Regras do Twitter**. [s.l.] 2021. Disponível em: https://help.twitter.com/pt/rules-and-policies/twitter-rules Acesso em: 18 mar. 2021.

TWITTER. **Política de informações enganosas sobre a COVID-19**. [s.l.] 2021a. Disponível em: https://help.twitter.com/pt/rules-and-policies/medical-misinformation-policy Acesso em: 18 mar. 2021.

TWITTER. **Política de mídia sensível**. [s.l.] 2019. Disponível em: https://help.twitter.com/pt/rules-and-policies/media-policy Acesso em: 18 mar. 2021.

UNIÃO EUROPEIA. Comissão Europeia. **The Digital Services Act**: ensuring a safe and accountable online environment, 2021. Disponível em: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en. Acesso em: 02 mar. 2021.