www.revistaft.com.br

Revista FT | ISSN 1678-0817 | v. 30, n. 158, 2026 | DOI: 10.69849/vvsn6v72

Liberdade de expressão *versus* discursos de ódio nas redes sociais

Freedom of expression versus hate speech on social media

Gabriel Rodrigues de Souza[1]
Marina Teodoro[2]
Rafael Rodrigues Alves[3]

Resumo

Este artigo aborda a relação entre liberdade de expressão e o enfrentamento do discurso de ódio nas redes sociais, no contexto da sociedade em rede e da atuação das plataformas digitais na circulação de conteúdos. O problema da pesquisa consiste em compreender como conciliar o direito à liberdade de expressão com o combate ao discurso de ódio no ambiente digital, considerando a moderação realizada por plataformas e seus impactos na visibilidade das informações. O objetivo geral consiste em analisar critérios jurídicos e desafios para o equilíbrio entre esses direitos fundamentais. Os objetivos específicos consistem em analisar a evolução histórica da liberdade de expressão, distinguir conceitualmente o discurso de ódio, investigar seus impactos sociais e psicológicos e propor parâmetros jurídicos de moderação e educação digital. A metodologia utilizada é a revisão de literatura, de caráter qualitativo e descritivo, com levantamento de doutrina, artigos científicos, legislação e jurisprudência em bases como CAPES, SciELO e Google Acadêmico, além de repositórios institucionais. Ao final, conclui-se que liberdade de expressão e dignidade humana convivem em tensão no ambiente digital, sendo necessária a aplicação da proporcionalidade para diferenciar crítica e discurso de ódio, além do fortalecimento da responsabilidade das plataformas, da transparência e da educação digital como mecanismos de proteção democrática.

[1] Universidade Evangélica- Campus Ceres- Goiás- Brasil. ORCID: https://orcid.org/0009-0006-9368-6794
[2] Universidade Evangélica- Campus Ceres- Goiás- Brasil. ORCID: https://orcid.org/0009-0004-4001-2900
[3] Universidade Evangélica- Campus Ceres- Goiás- Brasil. ORCID: https://orcid.org/0009-0002-6378-9208

Palavras-Chave: Discurso De Ódio. Liberdade De Expressão. Redes Sociais.

Abstract

This article addresses the relationship between freedom of expression and the confrontation of hate speech on social media, within the context of the network society and the role of digital platforms in the circulation of content. The research problem consists of understanding how to reconcile the right to freedom of expression with the fight against hate speech in the digital environment, considering moderation carried out by platforms and its impacts on the visibility of information. The general objective is to analyze legal criteria and challenges for balancing these fundamental rights. The specific objectives include analyzing the historical evolution of freedom of expression, conceptually distinguishing hate speech, investigating its social and psychological impacts, and proposing legal parameters for moderation and digital education. The methodology used is a literature review, with a qualitative and descriptive approach, based on the examination of legal doctrine, scientific articles, legislation, and case law from databases such as CAPES, SciELO, and Google Scholar, as well as institutional repositories. In conclusion, it is found that freedom of expression and human dignity coexist in tension in the digital environment, requiring the application of proportionality to distinguish between criticism and hate speech, as well as the strengthening of platform accountability, transparency, and digital education as mechanisms for democratic protection.
Keywords: Hate Speech. Freedom of Expression. Social Media.

1. Introdução

A comunicação humana evolui da fala e da escrita para o ambiente digital alterando as formas de interação e produção de conhecimento. Nesse contexto, a informação deixa de seguir um fluxo linear e passa a circular em múltiplas direções, com maior participação dos usuários. Entretanto, a liberdade de expressão, historicamente ligada à ampliação do debate público, passa a operar sob a mediação de plataformas digitais, as quais definem critérios de visibilidade por meio de regras próprias e sistemas automatizados, influenciando o alcance dos conteúdos.

No campo jurídico, busca-se permitir a discordância sem abrir espaço para ataques que afastem certos grupos do debate, por isso, surgiu então a pergunta: como conciliar o direito à liberdade de expressão com o combate ao discurso de ódio nas redes sociais?

A partir dessa questão, este estudo definiu como objetivo geral analisar critérios jurídicos e dificuldades que envolvem o equilíbrio entre liberdade de expressão e o enfrentamento do discurso de ódio nas redes. Para que fosse possível alcançar as respostas pretendidas e seguir a fundamentação teórica proposta, foram elencados objetivos específicos que perpassaram, primeiramente, pela descrição da evolução histórica da liberdade de expressão e sua distinção conceitual em relação ao discurso de ódio.

Em um segundo momento, a pesquisa propôs examinar os impactos sociais e psicológicos causados pela propagação do discurso de ódio nas plataformas digitais, compreendendo como essas manifestações afetam a saúde mental e a participação das vítimas na esfera pública. E, como último objetivo, o trabalho buscou propor parâmetros jurídicos de moderação e medidas de educação digital para a proteção da dignidade humana nas redes sociais, consolidando uma proposta de governança fundamentada na proporcionalidade e na transparência.

2. Revisão da Literatura

2.1 Uma nova realidade na comunicação: liberdade de expressão e discurso de ódio

A história da comunicação é marcada por transformações que alteraram a organização social e a produção do conhecimento. Inicialmente, a oralidade era o principal meio de transmissão cultural, baseada na memória coletiva e na repetição.

Segundo Walter J. Ong, nas culturas orais primárias, a redundância era necessária para preservar informações diante da ausência de registros escritos (Ong, 2008). Com a escrita, tornou-se possível registrar e conservar o conhecimento para além do tempo e do espaço, favorecendo o pensamento abstrato e a análise lógica.

No século XV, a imprensa de Gutenberg ampliou a circulação de textos e consolidou a leitura individual. Marshall McLuhan afirma que esse processo fortaleceu uma cultura visual e linear, além de impulsionar movimentos como a Reforma Protestante e o Iluminismo (McLuhan, 1972).

Com a Revolução Industrial, surgiram os meios de comunicação de massa, como rádio, televisão e cinema. Para McLuhan (1964), esses meios funcionam como extensões dos sentidos humanos, influenciando diretamente a organização social e consolidando a ideia de “aldeia global”.

Ong (2008) também desenvolve o conceito de “oralidade secundária”, relacionado aos meios eletrônicos que retomam características da oralidade mediadas pela tecnologia.

Com a internet, a comunicação passou a ser marcada pela interatividade e pela produção colaborativa. Manuel Castells define esse contexto como “sociedade em rede”, enquanto Henry Jenkins destaca a cultura da convergência e a participação ativa dos usuários (Castells, 2009; Jenkins, 2009).

André Lemos e Pierre Lévy (2010) associam esse cenário à inteligência coletiva. Paralelamente, surgem debates sobre veracidade das informações e moderação de conteúdos, enfrentados pelo Marco Civil da Internet e por diretrizes da Unesco (Unesco, 2023).

A liberdade de expressão consolidou-se como um dos pilares das democracias modernas a partir do constitucionalismo, das revoluções liberais e da proteção internacional dos direitos humanos. John Milton, em Areopagitica, defendeu a livre circulação de ideias e criticou a censura (Milton, 2004). Já John Stuart Mill, em Sobre a Liberdade, sustentou que a expressão deve ser livre, salvo quando causar dano a terceiros (Mill, 1991).

No plano jurídico, esse direito foi incorporado pela Declaração dos Direitos do Homem e do Cidadão e consolidado internacionalmente pela Declaração Universal dos Direitos Humanos, pelo Pacto Internacional sobre Direitos Civis e Políticos e pela Convenção Americana de Direitos Humanos, que vedam a censura prévia e admitem restrições legais proporcionais.

No Brasil, a Constituição Federal de 1988 garantiu amplamente a liberdade de expressão e proibiu a censura prévia (Brasil, 1988). O Supremo Tribunal Federal reforçou essa proteção na ADPF 130, ao declarar incompatível a antiga Lei de Imprensa com a Constituição (Brasil, 2009). Na doutrina, Daniel Sarmento (2016) aponta a liberdade de expressão como condição da democracia, enquanto Luís Roberto Barroso (2004) destaca sua dimensão individual e coletiva.

O Marco Civil estabelece princípios como neutralidade de rede, privacidade e liberdade de expressão, além de regras sobre responsabilidade de plataformas (Brasil, 2014). Já a LGPD regula o tratamento de dados pessoais e fortalece a proteção da privacidade no ambiente digital (Brasil, 2018). No plano internacional, documentos da Organização das Nações Unidas defendem

que restrições à expressão devem observar a legalidade, necessidade e proporcionalidade (ONU, 2011; Kaye, 2018).

No Brasil, o STF, ao julgar os Temas 987 e 533 em 2025, redefiniu parâmetros de responsabilidade das plataformas digitais, afastando a exigência exclusiva de ordem judicial em hipóteses de ilicitude evidente (STF, 2025). Isso impacta diretamente a moderação de conteúdos e a governança algorítmica (Souza; Lemos, 2016).

Dados do Cetic.br mostram ampla digitalização da população brasileira, ampliando o poder de curadoria das plataformas (CGI.br, 2025). Nesse contexto, o Tribunal Superior Eleitoral também desenvolve estratégias de enfrentamento à desinformação (TSE, 2024).

A doutrina destaca a necessidade de equilíbrio entre liberdade de expressão e proteção de outros direitos fundamentais. Daniel Sarmento (2006) reconhece limites legítimos à expressão em casos de discurso de ódio, enquanto Virgílio Afonso da Silva (2002) defende a proporcionalidade como critério de ponderação.

A delimitação do discurso de ódio é necessária para distingui-lo de críticas ou opiniões impopulares. Juridicamente, trata-se de manifestações dirigidas a grupos identificáveis por raça, etnia, religião, nacionalidade ou outras características, com potencial de estimular discriminação, hostilidade ou violência (OEA, 1969; ONU, 1966).

No Brasil, essa compreensão deve ser compatibilizada com a Constituição Federal de 1988, que protege a liberdade de expressão, mas também assegura dignidade, igualdade e combate ao racismo. Nesse sentido, o STF, no Caso Ellwanger, afastou a proteção constitucional de discursos antissemitas (Brasil, 2003).

O discurso de ódio diferencia-se da mera ofensa porque busca inferiorizar ou excluir grupos em razão de sua identidade, e não apenas criticar ideias ou comportamentos. Por isso, o direito internacional admite restrições proporcionais à liberdade de expressão para impedir práticas discriminatórias (OEA, 1969; ONU, 2011).

2.2 O discurso de ódio no ambiente digital: impactos sociais, psicológicos e suas interconexões

Os discursos de ódio nas redes sociais ultrapassam o ambiente digital ao afetar a convivência democrática, a coesão social e a proteção de grupos vulneráveis, ao reforçarem

hierarquias históricas e transformarem plataformas em espaços de disputa simbólica marcados por ofensas e estigmatização (Quadrado; Ferreira, 2020).

No ambiente digital, o discurso de ódio também atua como mecanismo de exclusão, ao influenciar quem pode participar do debate e em quais condições. A arquitetura das plataformas favorece agrupamentos por afinidade, expondo determinados grupos à hostilidade constante e reforçando desigualdades já existentes na sociedade offline, especialmente contra pessoas negras, mulheres, LGBTQIA+ e minorias religiosas (Jesus; Nobre, 2024).

Além disso, a lógica algorítmica das plataformas privilegia conteúdos polarizadores e emocionais, o que amplia a visibilidade de mensagens agressivas e discriminatórias, incentivando ataques e reduzindo a qualidade do debate público (Trindade, 2022).

Do ponto de vista institucional, o discurso de ódio compromete a esfera pública democrática ao restringir a participação de grupos vulneráveis, utilizando a própria liberdade de expressão como instrumento de intimidação e exclusão (Santos; Bueno, 2024). A ONU também destaca que essas práticas enfraquecem a confiança social e transformam o debate público em um ambiente de hostilidade constante (Nações Unidas, 2019).

No Brasil, o Estado reconhece a relação entre discurso de ódio e fenômenos como violência, racismo, xenofobia, misoginia e ataques antidemocráticos. O relatório oficial aponta que essas manifestações integram redes de mobilização que podem resultar em agressões físicas e ameaças a comunidades inteiras, evidenciando um problema social, e não apenas jurídico ou tecnológico (Brasil, 2023).

A literatura também caracteriza o discurso de ódio como violência simbólica que desumaniza grupos e prejudica a qualidade do debate público, gerando medo, desconfiança e enfraquecimento da ideia de igualdade de direitos (Batista; Ramos, 2024).

Esse fenômeno está associado à polarização política e à radicalização, ao transformar adversários em inimigos e reduzir o espaço para diálogo e consenso, o que fragiliza a convivência democrática (Quadrado; Ferreira, 2020; Batista; Ramos, 2024).

Os impactos do discurso de ódio são desigualmente distribuídos, atingindo principalmente grupos historicamente marginalizados, como pessoas negras, indígenas e LGBTQIA+, o que aprofunda vulnerabilidades e restringe sua participação na esfera pública (Brasil, 2023).

Em muitos casos, essas violências levam ao afastamento das vítimas do debate público, com restrição de perfis, exclusão de interações ou abandono das plataformas, configurando uma forma de expulsão simbólica do espaço digital (Brasil, 2023).

Estudos indicam que esses impactos também atingem a saúde mental, provocando sentimentos de humilhação, medo e isolamento, especialmente entre crianças e adolescentes expostos a ambientes hostis nas redes sociais (Cavalcante et al., 2024).

Do ponto de vista psicológico, o discurso de ódio nas redes sociais é entendido como violência simbólica que atinge a dignidade, a autoimagem e o sentimento de pertencimento das vítimas, gerando sofrimento emocional e sendo fator de risco para adoecimento mental em contextos de uso intenso das mídias digitais (Freitas et al., 2023; Batista; Ramos, 2024).

A exposição contínua a conteúdos hostis é frequentemente vivenciada como desqualificação pessoal, provocando vergonha, humilhação, medo e impotência. Em crianças e adolescentes, comentários depreciativos e estereótipos negativos contribuem para mal-estar persistente e impacto na percepção de si e na relação com o próprio corpo (Batista; Ramos, 2024; Taboga; Santos Junior, 2021).

A literatura aproxima o discurso de ódio do cyberbullying, sobretudo quando há repetição de ataques ligados a características pessoais como raça, gênero ou orientação sexual. Estudos indicam aumento de ansiedade, tristeza persistente, irritabilidade e ideação auto lesiva entre jovens expostos a essas agressões (Cavalcante et al., 2024; Taboga; Santos Junior, 2021).

Além disso, ambientes digitais marcados por hostilidade se associam a pior bem-estar emocional, incluindo solidão, distúrbios de sono, desinteresse por atividades e queda no desempenho escolar, podendo contribuir para quadros mais duradouros de sofrimento psíquico (Cavalcante et al., 2024; Brasil, 2024).

O discurso de ódio também afeta a autoimagem, especialmente em adolescentes, sendo relacionado a crises de ansiedade, autodepreciação e vergonha intensa, com maior impacto em meninas. Sentimentos de inferioridade se intensificam em contextos de comparação social e exclusão digital (Batista; Ramos, 2024; Taboga; Santos Junior, 2021).

A exposição repetida a mensagens ofensivas pode gerar baixa autoestima, sintomas depressivos, dificuldades de concentração e isolamento social, além de levar à evitação de espaços digitais e presenciais por medo de novas agressões (Cavalcante et al., 2024; Taboga; Santos Junior, 2021).

Outro efeito relevante é o impacto vicário, em que indivíduos do mesmo grupo das vítimas também sofrem sofrimento psicológico ao serem expostos a discursos de ódio, desenvolvendo sensação de ameaça, ansiedade e retração da participação pública (Batista; Ramos, 2024; Freitas et al., 2023).

Relatórios indicam ainda que o medo de retaliação e estigmatização dificulta a denúncia e a busca por ajuda, contribuindo para a manutenção da violência e agravamento dos sintomas, especialmente quando há falhas de apoio familiar ou escolar (Brasil, 2023; Brasil, 2024).

Organismos internacionais reconhecem o discurso de ódio online como fator de risco relevante à saúde mental, especialmente entre jovens, sendo comparável, em termos de impacto psíquico, a agressões presenciais, exigindo estratégias de prevenção e cuidado (Nações Unidas, 2019; Brasil, 2023).

Assim, os impactos psicológicos incluem ansiedade, depressão, vergonha, ideação auto lesiva e retração social, estando diretamente ligados às condições de exclusão e desumanização no ambiente digital, o que se conecta aos impactos sociais já analisados (Taboga; Santos Junior, 2021; Freitas et al., 2023).

Ademais, a literatura recente aponta que os impactos sociais e psicológicos do discurso de ódio nas redes sociais formam um circuito integrado, no qual práticas discriminatórias e sofrimento psíquico se retroalimentam (Batista; Ramos, 2024). Nesse contexto, as agressões simbólicas não são eventos isolados, mas inserem-se em dinâmicas coletivas de exclusão que influenciam a percepção de si e do mundo social (Freitas et al., 2023).

Estudos indicam que mensagens hostis degradam a esfera pública, fragilizam a confiança entre grupos e criam um ambiente de ameaça constante, associado a estresse, medo e hipervigilância entre os alvos (Quadrado; Ferreira, 2020). Assim, o uso das redes deixa de ser espaço de expressão e passa a ser percebido como fonte de insegurança e ansiedade (Batista; Ramos, 2024).

Nesse cenário, o discurso de ódio naturaliza a exclusão e a humilhação pública de minorias, reforçando desigualdades e produzindo sentimentos de vergonha, impotência e desvalia, que podem levar à autoexclusão do debate público (Batista; Ramos, 2024; Santos; Bueno, 2024).

Pesquisas sobre bolhas sociais mostram que o afastamento das vítimas favorece ambientes mais homogêneos, nos quais discursos de ódio circulam com menor contestação, o que reforça visões excludentes e aumenta a recorrência de novas agressões (Jesus; Nobre, 2024; Freitas et al., 2023).

Revisões de literatura indicam que a exposição contínua a interações hostis está associada à ansiedade, tristeza e insegurança em crianças e adolescentes, com reflexos na vida escolar,

familiar e social, afetando concentração, motivação e vínculos de confiança (Cavalcante et al., 2024; Taboga; Santos Junior, 2021).

Dados nacionais apontam que a incitação à violência na internet gera medo real de agressões e perseguições, intensificando estados de hipervigilância e desconfiança entre grupos vulneráveis (Brasil, 2024).

Em perspectiva internacional, a ONU destaca que o discurso de ódio ameaça a dignidade humana e a coesão social, sendo seus efeitos psicológicos inseparáveis de dinâmicas estruturais de violência e exclusão (Nações Unidas, 2019; Brasil, 2023).

Relatórios nacionais também indicam que o medo de retaliação reduz a busca por canais de denúncia e participação política, limitando o exercício de direitos e enfraquecendo a pluralidade no espaço público (Brasil, 2023; Santos; Bueno, 2024).

Além disso, a perda de autoestima, o sentimento de não pertencimento e a insegurança podem agravar quadros de sofrimento psíquico e intensificar o isolamento social, especialmente em contextos marcados por desigualdade e vulnerabilidade (Cavalcante et al., 2024; Batista; Ramos, 2024).

Diante disso, políticas de enfrentamento ao discurso de ódio defendem uma abordagem integrada que combine educação em direitos humanos, regulação de plataformas e cuidado em saúde mental, buscando reduzir simultaneamente a violência simbólica e seus efeitos psicológicos (Nações Unidas, 2019; Brasil, 2023).

Deste modo, os impactos sociais e psicológicos do discurso de ódio não são periféricos, mas expressam violações diretas à dignidade humana e à participação democrática. Quando a expressão passa a produzir medo, exclusão e silenciamento, torna-se necessário discutir limites jurídicos proporcionais à liberdade de expressão (Nações Unidas, 2019; Brasil, 2023).

2.3 Liberdade de expressão e discurso de ódio no ambiente digital: colisão de direitos fundamentais, desafios regulatórios e parâmetros de responsabilização

No ambiente digital, o equilíbrio entre liberdade de expressão e combate ao discurso de ódio deixou de ser uma questão resolvida apenas pela oposição entre fala livre e censura. Nas redes sociais, a circulação de conteúdos ocorre em estruturas privadas de comunicação em massa, nas quais plataformas definem regras, organizam visibilidade, impulsionam conteúdos, reduzem alcance e removem publicações com base em sistemas técnicos e critérios internos. Por isso, a

controvérsia jurídica contemporânea não diz respeito apenas ao que pode ou não ser dito, mas também a como a fala é distribuída, por quem ela é administrada e com quais salvaguardas se controla a intervenção sobre o debate público (Gorwa; Binns; Katzenbach, 2020; Douek, 2022).

No plano constitucional e internacional, essa discussão exige reconhecer que a liberdade de expressão possui papel central na democracia, mas não se reveste de caráter absoluto. No Brasil, o Supremo Tribunal Federal firmou, na ADPF 130, entendimento fortemente contrário à censura prévia, reconhecendo a posição de destaque da liberdade de expressão e de imprensa na ordem constitucional.

Ao mesmo tempo, no julgamento do HC 82.424, o caso Ellwanger, a Corte deixou claro que manifestações racistas não podem ser legitimadas sob o rótulo de exercício regular da liberdade de expressão. Em sentido convergente, o artigo 13 da Convenção Americana sobre Direitos Humanos e o Comentário Geral n. 34 do Comitê de Direitos Humanos da ONU admitem restrições à expressão apenas quando previstas em lei, voltadas à proteção de direitos ou bens constitucionalmente relevantes e compatíveis com exigências estritas de necessidade e proporcionalidade (Brasil, 2009; 2003; OEA, 1969; ONU, 2011).

É nesse ponto que a proporcionalidade se torna um critério essencial de análise, porque, sob o aspecto da adequação, a imposição de limites jurídicos a manifestações que configuram discurso de ódio se mostra apta a proteger a dignidade humana, a igualdade material e a participação segura de pessoas e grupos vulnerabilizados no espaço público, especialmente quando a fala promove inferiorização, desumanização ou estímulo à discriminação, à hostilidade ou à violência, em consonância com parâmetros internacionais recentes que reconhecem o discurso de ódio como ameaça direta aos direitos humanos, à democracia e à convivência plural e, por isso, exigem respostas juridicamente estruturadas, e não meramente intuitivas (Conselho da Europa, 2022).

No entanto, reconhecer que a intervenção pode ser adequada não resolve, por si só, o problema da necessidade, porque a noção de discurso de ódio costuma ser usada de forma ampla e isso pode gerar tanto omissão quanto excesso, já que definições vagas abrem espaço para arbitrariedade e censura indevida, enquanto definições rígidas demais não acompanham formas atuais de agressão, como campanhas coordenadas, linguagem codificada, memes e assédio reiterado, razão pela qual a restrição jurídica só se justifica diante de manifestações que, observadas em seu alvo, conteúdo, contexto e potencial lesivo, deixam de ser mera crítica

ofensiva e passam a funcionar como mecanismo de exclusão e inferiorização de pessoas ou grupos, de modo que o

O desafio não está em reprimir toda fala chocante, mas em construir critérios claros o suficiente para conter o ódio sem transformar dissenso, ironia, crítica política ou opinião impopular em ilícito (OEA, 1969; ONU, 2011; Conselho da Europa, 2022).

A proporcionalidade em sentido estrito pede um cuidado ainda maior porque, no ambiente digital, regular a circulação da fala não significa apenas remover conteúdos, mas também lidar com medidas como desindexação, desmonetização, limitação de recomendações, rebaixamento algorítmico e outras formas de redução de visibilidade que, embora possam ser úteis em casos graves, também podem gerar falta de clareza, erros e autocensura quando não vêm acompanhadas de transparência, aviso ao usuário, justificativa mínima e possibilidade real de recurso, de modo que o problema deixa de ser apenas saber se o conteúdo foi retirado e passa a incluir a necessidade de garantir que a pessoa compreenda o que aconteceu, por qual motivo e de que forma pode contestar a decisão, o que explica por que propostas regulatórias mais recentes insistem em devido processo informacional e em padrões mínimos de accountability (Santa Clara Principles, 2021; Leerssen, 2023; União Europeia, 2022).

Outro desafio relevante decorre da própria estrutura técnica e organizacional da moderação em larga escala. A triagem de conteúdos em plataformas globais depende de arranjos híbridos, que combinam revisão humana, automação, priorização por risco e procedimentos internos fragmentados. A literatura mostra que sistemas automatizados enfrentam dificuldades relevantes para lidar com contexto, ironia, ambiguidade, idioma, reapropriação discursiva e códigos culturais, o que impede tratar a tecnologia como solução neutra ou suficiente para definir o que constitui discurso de ódio em todos os casos. Além disso, a qualidade das decisões também depende das condições concretas de trabalho dos moderadores humanos, cuja atuação ocorre, muitas vezes, sob pressão, alta exposição a conteúdos perturbadores e exigência de produtividade em escala industrial (Gorwa; Binns; Katzenbach, 2020; Gillespie, 2018; Roberts, 2019).

Esse quadro mostra que a resposta jurídica não pode partir da ideia irreal de decisões perfeitas nem da expectativa de que as plataformas ajam de forma naturalmente neutra, assim como também não basta atribuir a elas uma responsabilidade genérica e total, como se fossem árbitras absolutas da esfera pública, já que o verdadeiro desafio está em definir deveres compatíveis com os direitos fundamentais, com regras mais claras, justificativas acessíveis, preservação de evidências, canais de denúncia, possibilidades de recurso e uma divisão funcional

de responsabilidades entre detecção, decisão, aplicação, revisão e auditoria, de modo que, mais do que exigir remoção automática, o essencial é estruturar sistemas de governança capazes de reduzir arbitrariedades, ampliar o controle público e permitir respostas proporcionais diante de riscos reais de discriminação e violência (Douek, 2022; Mulligan; Bamberger, 2021).

Assim, o principal desafio deste debate é definir parâmetros normativos capazes de preservar o núcleo democrático da liberdade de expressão sem permitir que ela seja instrumentalizada como escudo para práticas de humilhação, exclusão e incitação contra grupos vulnerabilizados. É a partir dessa colisão, e da necessidade de resolvê-la com base em adequação, necessidade e proporcionalidade em sentido estrito, que se torna possível avançar para os critérios jurídicos de distinção entre crítica ofensiva e discurso de ódio, bem como para a delimitação da responsabilidade das plataformas no tratamento dessas manifestações.

O enfrentamento do discurso de ódio nas redes sociais tende a ser mais consistente quando combina resposta jurídica com prevenção social, razão pela qual a educação digital deve ser compreendida como estratégia complementar de redução de danos e fortalecimento da convivência democrática. Isso significa reconhecer que o problema não se resolve apenas pela remoção de conteúdos ou pela imposição de sanções, mas também pela formação de usuários capazes de identificar práticas discriminatórias, compreender o funcionamento das plataformas e participar do debate público com responsabilidade. Nessa perspectiva, a educação digital não se confunde com censura nem substitui a necessidade de critérios jurídicos claros, mas contribui para reduzir vulnerabilidades, ampliar a capacidade crítica dos sujeitos e enfraquecer a normalização social do discurso de ódio (UNESCO, 2023; Nações Unidas, 2023).

Em uma compreensão mais ampla, a educação digital se aproxima do campo da media and information literacy porque envolve habilidades para acessar, avaliar, interpretar, produzir e compartilhar informações de forma crítica, ética e situada no contexto em que circulam, e, quando aplicada ao problema do discurso de ódio, passa a incluir a capacidade de reconhecer linguagem desumanizante, perceber dinâmicas de viralização e recomendação, distinguir a crítica legítima de ataques dirigidos a pessoas ou grupos vulnerabilizados e adotar formas responsáveis de reação, como denunciar, buscar apoio institucional e responder sem ampliar o dano, de modo que seu papel não se limita a ensinar regras de uso seguro da internet, mas a preparar o usuário para compreender que a liberdade de expressão, embora essencial ao pluralismo democrático, não pode ser usada para legitimar práticas de humilhação, exclusão e intimidação que ferem a igual dignidade no espaço digital (UNESCO, 2023).

A relevância dessa formação não é apenas teórica. Pesquisa desenvolvida por Obermaier e Schmuck com adolescentes e jovens adultos mostrou que níveis mais elevados de digital media literacy estiveram associados a menor propensão à vitimização por ódio online em diferentes categorias e a maior probabilidade de pertencimento a perfis de baixa exposição ao fenômeno. O dado é importante porque sugere que competências de navegação crítica, percepção de risco e leitura qualificada do ambiente digital funcionam como fator protetivo em contextos marcados por assimetrias informacionais, polarização e exposição recorrente a conteúdos hostis. Assim, a educação digital deixa de ser tratada como simples conteúdo escolar acessório e passa a ser compreendida como elemento de proteção social e de fortalecimento da autonomia dos usuários diante de ecossistemas comunicacionais complexos (Obermaier; Schmuck, 2022).

No campo pedagógico, a prevenção também depende de ações contínuas e bem estruturadas, e estudo recente sobre o programa HateLess mostrou que intervenções educativas voltadas a adolescentes podem reduzir tanto a prática quanto a vitimização por discurso de ódio on-line, ao mesmo tempo em que fortalecem respostas discursivas de enfrentamento, empatia e confiança para agir diante de conteúdos discriminatórios, o que reforça a importância de programas que articulem alfabetização informacional, educação em direitos humanos, capacidades argumentativas, reconhecimento do dano e habilidades socioemocionais, evitando abordagens superficiais baseadas apenas em alertas genéricos sobre bom comportamento na internet, já que a prevenção tende a ser mais eficaz quando prepara as pessoas para reconhecer o problema, responder de forma proporcional e compreender os efeitos concretos da violência simbólica sobre indivíduos e grupos (Wachs et al., 2024; UNESCO, 2023).

A conscientização social precisa ser pensada de forma articulada entre escola, família, mídia, sociedade civil e plataformas, já que a vivência digital não acontece só no espaço escolar, o que exige políticas de formação de professores, materiais pedagógicos, campanhas públicas e iniciativas voltadas aos usuários que dialoguem com situações reais de exposição ao ódio on-line e orientem sobre identificação de riscos, preservação de evidências, canais de denúncia, busca de apoio e formas de reação que não ampliem a violência, de modo que a educação digital contribua para sustentar um ambiente público mais plural e menos hostil, reduzindo a aceitação do ódio como entretenimento, brincadeira ou manifestação supostamente natural da liberdade, embora seja importante reconhecer que ela não substitui a definição de critérios jurídicos para distinguir crítica ofensiva de discurso de ódio nem afasta a necessidade de delimitar a responsabilidade das plataformas, tendo sua principal contribuição na redução da dependência exclusiva de respostas

repressivas e no fortalecimento, no plano cultural e formativo, de condições mais adequadas para o exercício responsável da liberdade de expressão (Nações Unidas, 2023; UNESCO, 2023).

As propostas legislativas e regulatórias voltadas ao enfrentamento do discurso de ódio nas redes sociais precisam partir de uma premissa básica: o problema não se resolve com ordens genéricas de remoção, porque a circulação de conteúdos ofensivos e discriminatórios depende de sistemas de recomendação, impulsionamento, despriorização, automação e gestão de riscos que atuam em larga escala. Por isso, a resposta jurídica mais compatível com a Constituição não é a censura prévia nem a liberdade irrestrita, mas a construção de limites proporcionais, controláveis e fundamentados, capazes de proteger o dissenso legítimo sem permitir que a liberdade de expressão seja utilizada como escudo para práticas de humilhação, inferiorização e exclusão de pessoas ou grupos vulnerabilizados (Brasil, 2009; 2003; OEA, 1969; ONU, 2011; Douek, 2022).

Um primeiro eixo regulatório consiste em definir, com maior precisão, o que distingue a crítica ofensiva do discurso de ódio. Nem toda manifestação agressiva, injusta ou moralmente reprovável pode ser tratada como ilícita, sob pena de esvaziar a liberdade de expressão e empobrecer o debate público. A intervenção jurídica mais intensa se justifica quando a manifestação, considerada em seu alvo, conteúdo, contexto e potencial lesivo, deixa de incidir apenas sobre ideias, opiniões ou condutas e passa a inferiorizar, desumanizar, estigmatizar ou estimular discriminação, hostilidade ou violência contra pessoas ou coletividades identificáveis. Esse critério é compatível com o entendimento do Supremo Tribunal Federal no caso Ellwanger, com a Convenção Americana sobre Direitos Humanos, com o Pacto Internacional sobre Direitos Civis e Políticos e com a Recommendation CM/Rec(2022)16 do Conselho da Europa, todos convergentes no sentido de que a proteção da expressão não alcança com a mesma intensidade manifestações que atentem contra a igual dignidade humana e favoreçam práticas discriminatórias (Brasil, 2003; OEA, 1969; ONU, 2011; Conselho da Europa, 2022).

Essa diferenciação exige que o intérprete observe, de maneira articulada, pelo menos quatro elementos. O primeiro é o alvo da manifestação, porque a crítica ofensiva, em regra, recai sobre ideias, comportamentos, posicionamentos políticos, instituições ou figuras públicas, ainda que com linguagem ácida, exagerada ou socialmente censurável, ao passo que o discurso de ódio se dirige a pessoas ou grupos identificáveis por características como raça, origem, religião, gênero ou outras condições de pertencimento, atribuindo-lhes inferioridade ou periculosidade. O segundo é o conteúdo da mensagem, já que não basta haver rudeza ou grosseria para que se justifique a intervenção mais intensa do direito. O que torna a manifestação mais grave é seu teor

de desumanização, estigmatização, exclusão ou associação degradante, sobretudo quando ela apresenta determinado grupo como indigno de igual respeito, convivência ou proteção. O terceiro elemento é o contexto, pois a mesma expressão não pode ser analisada de forma isolada quando integra campanha reiterada, ambiente hostil, histórico discriminatório, dinâmica de assédio coordenado ou cenário de vulnerabilidade acentuada do grupo atingido. O quarto é o potencial lesivo, entendido como a aptidão concreta da fala para estimular discriminação, hostilidade ou violência, ainda que não haja convocação explícita e imediata para agressão física. Esses vetores ajudam a evitar tanto a banalização do conceito de discurso de ódio quanto sua compressão excessiva, permitindo uma leitura mais compatível com a proteção reforçada da liberdade de expressão e, ao mesmo tempo, com a tutela da dignidade humana e da igualdade material (OEA, 1969; ONU, 1966; ONU, 2011; Conselho da Europa, 2022).

Sob essa perspectiva, o critério jurídico adequado não está em saber apenas se a fala ofendeu, chocou ou causou desconforto, mas em verificar se ela permanecer no campo do dissenso, ainda que duro, ou se passou a funcionar como forma de inferiorização e exclusão social, pois, em sociedades democráticas, opiniões impopulares, críticas severas, sátiras e até manifestações de mau gosto continuam a receber proteção relevante justamente para evitar que o Estado ou agentes privados silenciem discursos apenas por serem ásperos, incômodos ou desconfortáveis.

O tratamento jurídico muda quando a manifestação deixa de enfrentar ideias e passa a atingir a dignidade de pessoas ou grupos, transformando características de identidade em motivo de humilhação, ameaça ou exclusão simbólica do espaço público, razão pela qual a liberdade de expressão não pode servir de proteção ao racismo nem a mensagens que reforcem práticas discriminatórias incompatíveis com a ordem constitucional, de modo que o uso desses critérios não busca ampliar indevidamente o campo da proibição, mas reduzir arbitrariedades e oferecer base mais segura para decisões do Estado e das plataformas, concentrando a resposta jurídica em manifestações realmente lesivas à convivência democrática, sem transformar o direito em instrumento de censura ampla ou de punição da divergência legítima (Brasil, 2003; OEA, 1969; ONU, 1966; ONU, 2011; Conselho da Europa, 2022)

Um segundo eixo envolve a incorporação de garantias procedimentais na governança das plataformas. Como a moderação de conteúdo é exercida cotidianamente por agentes privados com impacto direto sobre a esfera pública, remoções, suspensões, desmonetizações e reduções de alcance não podem ocorrer sem regras claras, notificação compreensível, indicação do

fundamento da medida e possibilidade efetiva de recurso. A exigência de devido processo informacional ganha ainda mais relevância quando são utilizadas técnicas automatizadas ou medidas de visibilidade pouco perceptíveis para o usuário, como delisting e demotion, porque nesses casos o afetado pode sequer compreender que foi sancionado ou por qual motivo. Nesse ponto, a literatura sobre moderação e as referências regulatórias mais influentes convergem ao sustentar que transparência, motivação e revisão não são detalhes acessórios, más condições mínimas de legitimidade para a atuação das plataformas (Santa Clara Principles, 2021; Leerssen, 2023; Keller, 2022; Douek, 2022).

Um terceiro eixo diz respeito aos deveres de diligência e à gestão de riscos sistêmicos. A experiência europeia com o Digital Services Act reforçou a ideia de que a regulação não deve ficar limitada à responsabilização posterior de casos isolados, mas também exigir das plataformas medidas estruturais de avaliação, mitigação e transparência proporcionais ao porte e ao impacto do serviço. No tema do discurso de ódio, isso significa deslocar o foco do post individual para o desenho do sistema, com atenção a políticas internas, canais de denúncia, monitoramento de campanhas coordenadas, funcionamento de mecanismos de recomendação e incentivos que favoreçam a viralização de conteúdos extremos. Não se trata de exigir remoção perfeita nem de transformar provedores em censores gerais, mas de impor deveres de diligência demonstráveis quando a arquitetura da plataforma amplia riscos previsíveis de discriminação, assédio e intimidação coletiva (União Europeia, 2022; Douek, 2022; Gorwa; Binns; Katzenbach, 2020).

Um quarto eixo diz respeito à transparência e à prestação de contas em sentido concreto, e não apenas formal, porque relatórios genéricos não são suficientes para saber se a moderação realmente reduz a circulação do discurso de ódio ou apenas desloca seus efeitos, sendo necessário que existam informações minimamente acessíveis sobre as regras aplicadas, os tipos de denúncia, as medidas adotadas, as taxas de revisão, o funcionamento básico dos sistemas de recomendação e os impactos diferenciados sobre grupos vulneráveis, sempre com o cuidado de não violar a privacidade nem facilitar o uso malicioso das regras, de modo que a transparência não deve ser máxima e indiscriminada, mas suficiente para permitir controle público, pesquisa qualificada e fiscalização institucional sobre decisões que afetam direitos fundamentais no ambiente digital (Keller, 2022; Santa Clara Principles, 2021; União Europeia, 2022).

Um quinto eixo exige atenção especial à proteção de grupos vulnerabilizados e à prevenção de danos mais graves, especialmente quando houver ameaça, perseguição, assédio reiterado ou ataques dirigidos a crianças, adolescentes e minorias historicamente expostas à

discriminação. No Brasil, a Lei n. 14.532/2023 reforçou a tutela contra práticas racistas ao equiparar a injúria racial ao crime de racismo, e a Lei n. 14.811/2024 passou a prever medidas de proteção à criança e ao adolescente, incluindo previsão legal relativa ao bullying e ao cyberbullying. Esses marcos não autorizam restrição indiscriminada do debate público, mas indicam que o ordenamento brasileiro já reconhece a necessidade de respostas mais claras e prioritárias diante de práticas que, além de ofender, produzem exclusão, medo e comprometimento da dignidade humana em contextos de especial vulnerabilidade (Brasil, 2023; 2024).

Por fim, a solução mais adequada não está em concentrar toda a responsabilidade no Estado, nas plataformas ou nos usuários, mas em adotar um modelo de responsabilidades distribuídas. Cabe ao direito definir critérios materiais e procedimentais compatíveis com a liberdade de expressão; às plataformas, cumprir deveres de transparência, diligência, recurso e mitigação de riscos; e à sociedade, inclusive por meio da educação digital, fortalecer práticas de resposta crítica e rejeição pública ao ódio (Mulligan; Bamberger, 2021; Conselho da Europa, 2022; ONU, 2011).

3. Metodologia

A pesquisa adotou metodologia de revisão de literatura e análise documental, com abordagem qualitativa e descritiva, a partir de livros, artigos científicos, dissertações, legislações e precedentes jurisprudenciais. As fontes foram obtidas em bases como o Portal de Periódicos da CAPES, SciELO e Google Acadêmico, além de repositórios institucionais do Supremo Tribunal Federal, do Ministério dos Direitos Humanos e da Cidadania e da Organização das Nações Unidas. O recorte temporal considerou publicações e marcos normativos entre 2002 e 2026, incluindo obras clássicas e decisões recentes sobre o tema. Os materiais foram submetidos à análise crítica, visando sustentar a discussão sobre colisão de direitos fundamentais e regulação do ambiente digital.

4. Resultados e Discussão

Barroso (2004) e Sarmento (2016) reconhecem a liberdade de expressão como direito essencial ao pluralismo democrático, mas defendem sua compatibilização com a dignidade da

pessoa humana e a igualdade material. No mesmo sentido, a Convenção Americana de Direitos Humanos (OEA, 1969) e o Comentário Geral n. 34 da ONU (2011) admitem restrições proporcionais para a proteção de direitos fundamentais.

> A pesquisa demonstrou que o principal desafio jurídico não é restringir opiniões impopulares, mas diferenciar manifestações protegidas de discursos destinados à discriminação e à desumanização de grupos vulneráveis. O caso Ellwanger reforçou esse entendimento ao afastar a proteção constitucional de manifestações racistas.

Com base em Virgílio Afonso da Silva (2002), verificou-se que a proporcionalidade é o principal critério para solucionar conflitos entre liberdade de expressão e combate ao discurso de ódio. Embora limitações possam ser legítimas, restrições excessivas podem gerar censura e insegurança jurídica.

Os resultados também indicaram que as plataformas digitais ampliam a circulação de conteúdos hostis. Gorwa, Binns e Katzenbach (2020) e Trindade (2022) demonstram que algoritmos de recomendação favorecem conteúdos extremos e discriminatórios. Douek (2022) observa que as plataformas passaram a exercer funções semelhantes às de administradoras da esfera pública digital, sem controle democrático equivalente.

Nesse contexto, Leerssen (2023) e os Santa Clara Principles (2021) defendem maior transparência nos processos de moderação, com fundamentação, notificação e possibilidade de recurso.

A pesquisa evidenciou ainda que o discurso de ódio compromete a participação democrática e reforça a exclusão social. Quadrado e Ferreira (2020) e Santos e Bueno (2024) afirmam que essas práticas afastam grupos vulneráveis do debate público. Relatórios da ONU (2019) e do governo brasileiro (Brasil, 2023) relacionam o discurso de ódio à radicalização e ao enfraquecimento da convivência democrática.

Além dos impactos sociais, Batista e Ramos (2024) e Freitas et al. (2023) demonstram que o discurso de ódio produz medo, ansiedade, retração social e insegurança emocional. Cavalcante et al. (2024) e Taboga e Santos Junior (2021) identificaram efeitos semelhantes em crianças e adolescentes vítimas de ataques relacionados à raça, gênero e orientação sexual.

Jesus e Nobre (2024) acrescentam que bolhas sociais favorecem a repetição de discursos discriminatórios e reduzem a pluralidade do debate público.

A pesquisa identificou também limitações da moderação automatizada. Gorwa, Binns e Katzenbach (2020) apontam dificuldades na interpretação de contexto e linguagem codificada,

enquanto Roberts (2019) demonstra que a moderação em larga escala ainda depende de decisões humanas. Por isso, Douek (2022) e Mulligan e Bamberger (2021) defendem modelos de governança baseados em transparência e responsabilização proporcional.

Outro resultado relevante refere-se à educação digital como estratégia preventiva. UNESCO (2023), Nações Unidas (2023), Obermaier e Schmuck (2022) e Wachs et al. (2024) demonstram que alfabetização midiática, empatia e educação em direitos humanos reduzem a prática e a vitimização associadas ao discurso de ódio.

A análise das fontes também permitiu diferenciar crítica ofensiva e discurso de ódio. OEA (1969), ONU (2011) e Conselho da Europa (2022) afirmam que a liberdade de expressão não protege manifestações que incentivem discriminação, hostilidade ou violência contra grupos identificáveis. Keller (2022) e os Santa Clara Principles (2021) defendem critérios transparentes de moderação para evitar censura excessiva e insegurança jurídica.

Relatórios nacionais (Brasil, 2023; Brasil, 2024) demonstraram maior incidência de ataques contra pessoas negras, mulheres, população LGBTQIA+ e minorias religiosas, evidenciando que medidas exclusivamente repressivas são insuficientes sem prevenção, educação digital e transparência.

5. Conclusão

A presente pesquisa demonstrou que a liberdade de expressão permanece como elemento indispensável ao Estado Democrático de Direito e à preservação do pluralismo no ambiente digital. Contudo, verificou-se que sua proteção não pode ser interpretada de forma absoluta quando determinadas manifestações passam a operar como instrumentos de discriminação, exclusão e desumanização de pessoas ou grupos vulneráveis.

A análise histórica e jurídica permitiu observar que a consolidação da liberdade de expressão ocorreu vinculada à proteção do debate público e à rejeição da censura prévia, especialmente a partir da Constituição Federal de 1988 e dos sistemas internacionais de direitos humanos. Ao mesmo tempo, os referenciais doutrinários, jurisprudenciais e normativos analisados demonstraram que a própria ordem democrática impõe limites proporcionais quando a manifestação ultrapassa o campo do dissenso legítimo e passa a estimular hostilidade, discriminação ou violência.

Os resultados evidenciaram ainda que o ambiente digital ampliou a complexidade desse debate. As plataformas digitais deixaram de atuar apenas como espaços neutros de hospedagem e

passaram a influenciar diretamente a circulação de conteúdos por meio de algoritmos, sistemas de recomendação e mecanismos de moderação. Nesse contexto, verificou-se que o enfrentamento do discurso de ódio não depende exclusivamente da remoção de conteúdos, mas também da definição de critérios transparentes, proporcionais e compatíveis com as garantias fundamentais.

A pesquisa também demonstrou que o discurso de ódio produz consequências que ultrapassam a esfera individual, afetando a convivência democrática, a participação pública e a saúde mental das vítimas. Os estudos analisados identificaram impactos como medo, ansiedade, retração social e afastamento de grupos vulneráveis dos espaços de debate, evidenciando que a permanência de práticas discriminatórias compromete a igualdade material e reduz a pluralidade democrática.

Além disso, constatou-se que soluções exclusivamente repressivas ou automatizadas são insuficientes diante da complexidade das interações digitais. As limitações da moderação algorítmica, somadas ao risco de censura excessiva e à opacidade das decisões das plataformas, reforçam a necessidade de modelos de governança baseados em transparência, possibilidade de revisão, responsabilização proporcional e devido processo informacional.

A análise das fontes permitiu concluir, ainda, que a diferenciação entre crítica ofensiva e discurso de ódio depende da observação conjunta do contexto, do conteúdo, do alvo da manifestação e de seu potencial lesivo. Assim, verificou-se que manifestações dirigidas à inferiorização de grupos identificáveis não recebem o mesmo grau de proteção constitucional atribuído às críticas políticas, opiniões impopulares ou manifestações ásperas inerentes ao debate democrático.

Por fim, concluiu-se que a preservação da liberdade de expressão nas redes sociais exige equilíbrio entre proteção ao dissenso legítimo e tutela da dignidade humana. Nesse cenário, o enfrentamento do discurso de ódio demanda atuação conjunta do Estado, das plataformas digitais e da sociedade civil, por meio de critérios jurídicos proporcionais, mecanismos transparentes de moderação e estratégias de educação digital voltadas à redução da violência simbólica e ao fortalecimento de um espaço público plural e democrático.

Referências

BARROSO, Luís Roberto. Colisão entre liberdade de expressão e direitos da personalidade. **Revista de Direito Administrativo,** Rio de Janeiro, v. 235, p. 111-138, jan./mar. 2004.

BATISTA, Waleska Miguel; RAMOS, Gisele Motta. Discurso de ódio: descortinando as violências nas redes sociais. **Revista Direito e Práxis,** Rio de Janeiro, v. 15, n. 3, 2024.

BRASIL. **Constituição (1988).** Constituição da República Federativa do Brasil de 1988. Brasília, DF: Senado Federal, 1988.

BRASIL. **Lei n. 12.965, de 23 de abril de 2014.** Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil (Marco Civil da Internet). Diário Oficial da União: seção 1, Brasília, DF, 24 abr. 2014.

BRASIL. **Lei n. 13.709, de 14 de agosto de 2018.** Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018.

BRASIL. L**ei n. 14.532, de 11 de janeiro de 2023**. Altera a Lei n. 7.716, de 5 de janeiro de 1989 (Lei do Crime Racial), e o Decreto-Lei n. 2.848, de 7 de dezembro de 1940 (Código Penal), para tipificar como crime de racismo e injúria racial, prevê pena de suspensão de direito em caso de racismo praticado no contexto de atividade esportiva ou artística e prever pena para o racismo religioso e recreativo e para o praticado por funcionário público. Brasília, DF: Presidência da República, 2023.

BRASIL. **Lei n. 14.811, de 12 de janeiro de 2024.** Institui medidas de proteção à criança e ao adolescente contra a violência nos estabelecimentos educacionais ou similares, prevê a Política Nacional de Prevenção e Combate ao Abuso e Exploração Sexual da Criança e do Adolescente, altera o Decreto-Lei n. 2.848, de 7 de dezembro de 1940 (Código Penal), e outras normas. Brasília, DF: Presidência da República, 2024.

BRASIL. **Ministério dos Direitos Humanos e da Cidadania.** Incitação à violência contra a vida na internet lidera violações de direitos humanos com mais de 76 mil casos em cinco anos, aponta ObservaDH. Brasília, DF: Ministério dos Direitos Humanos e da Cidadania, 2024.

BRASIL. **Ministério dos Direitos Humanos e da Cidadania**. Relatório de recomendações para o enfrentamento do discurso de ódio e do extremismo no Brasil. Brasília, DF: Ministério dos Direitos Humanos e da Cidadania, 2023.

BRASIL. **Supremo Tribunal Federal. Arguição de Descumprimento de Preceito Fundamental n. 130/DF**. Relator: Min. Carlos Ayres Britto. Tribunal Pleno. Julgado em 30 abr. 2009. Diário da Justiça Eletrônico: Brasília, DF, 6 nov. 2009.

BRASIL. **Supremo Tribunal Federal. Habeas Corpus n. 82.424/RS**. Rel. Min. Moreira Alves, red. p/ o acórdão Min. Maurício Corrêa. Tribunal Pleno. Julgado em 17 set. 2003. Diário da Justiça: Brasília, DF, 19 mar. 2004.

BRASIL. Supremo Tribunal Federal. **Recurso Extraordinário com Agravo (ARE) n. 722.744/RS**. Relator: Min. Luís Roberto Barroso. Julgado em 17 dez. 2015. Diário da Justiça Eletrônico: Brasília, DF, 2015.

CASTELLS, Manuel. **A sociedade em rede.** 21. ed. São Paulo: Paz e Terra, 2009.

CAVALCANTE, Ester Oliveira et al. O impacto das redes sociais na saúde mental das crianças e adolescentes: um estudo de revisão da literatura. **RevistaFT, Goiânia,** v. 28, n. 135, 2024.

CGI.BR/NIC.BR. **TIC Domicílios 2024**: livro eletrônico. São Paulo: CGI.br/NIC.br, 2024.

COMITÊ DE DIREITOS HUMANOS DAS NAÇÕES UNIDAS. **General Comment No. 34**: Article 19: Freedoms of opinion and expression. CCPR/C/GC/34. Genebra: Organização das Nações Unidas, 2011. Publicado em 29 jul. 2011.

CONSELHO DA EUROPA. **Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech. Estrasburg**o: Council of Europe, 2022. Adotada em 20 maio 2022.

DOUEK, Evelyn. Content moderation as systems thinking. **Harvard Law Review,** Cambridge, v. 136, n. 2, p. 526-607, 2022.

FRANÇA. **Declaração dos Direitos do Homem e do Cidadão**. Paris, 1789.

FREITAS, Ana Luísa et al. Bases sociocognitivas do discurso de ódio online no Brasil. **Texto Livre: Linguagem e Tecnologia**, Campinas/Belo Horizonte, v. 16, e42662, 2023.

GILLESPIE, Tarleton. **Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media**. New Haven: Yale University Press, 2018.

GORWA, Robert; BINNS, Reuben; KATZENBACH, Christian. Algorithmic content moderation: technical and political challenges in the automation of platform governance. **Big Data & Society,** v. 7, n. 1, 2020.

JENKINS, Henry. **Cultura da convergência.** 2. ed. São Paulo: Aleph, 2009.

JESUS, Beatriz Pereira de; NOBRE, Thalita Lacerda. As bolhas sociais e o discurso de ódio nas redes sociais digitais. **Cadernos Zygmunt Bauman,** São Luís, v. 14, n. 36, 2024.

KAYE, David. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.** A/HRC/38/35. Genebra: Conselho de Direitos Humanos das Nações Unidas, 2018.

KELLER, Daphne; LEVY, Max. **Getting transparency right**. Lawfare, 11 jul. 2022.

LEERSSEN, Paddy. An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. **Computer Law & Security Review**, v. 48, art. 105790, 2023.

LEMOS, André; LÉVY, Pierre. **O futuro da internet:** em direção a uma ciberdemocracia planetária. São Paulo: Paulus, 2010.

MCLUHAN, Marshall. A **galáxia de Gutenberg:** a formação do homem tipográfico. São Paulo: Editora Nacional, 1972.

MCLUHAN, Marshall. **Os meios de comunicação como extensões do homem.** São Paulo: Cultrix, 1964.

MILL, John Stuart. **Sobre a liberdade**. Petrópolis: Vozes, 1991.

MILTON, John. **Areopagítica:** discurso pela liberdade de impressão ao Parlamento da Inglaterra. São Paulo: TopBooks, 2004.

MULLIGAN, Deirdre K.; BAMBERGER, Kenneth A. Allocating responsibility in content moderation: a functional framework. **Berkeley Technology Law Journal,** v. 36, n. 3, p. 10911172, 2021.

NAÇÕES UNIDAS. **Estratégia e Plano de Ação das Nações Unidas sobre o Discurso de Ódio**. Nova York: Organização das Nações Unidas, 2019.

OBERMAIER, Magdalena; SCHMUCK, Desirée. Youths as targets: factors of online hate speech victimization among adolescents and young adults. **Journal of Computer-Mediated Communication,** v. 27, n. 4, art. zmac012, 2022.

ONG, Walter J. **Oralidade e cultura escrita:** a tecnologia da palavra. Campinas: Papirus, 2008.

Organização das Nações Unidas. **Declaração Universal dos Direitos Humanos**. Paris, 1948.

Organização das Nações Unidas. **Our Common Agenda: Policy Brief 8 – Information Integrity on Digital Platforms.** New York: United Nations, 2023.

Organização das Nações Unidas. **Pacto Internacional sobre Direitos Civis e Políticos**. Nova Iorque, 1966.

Organização dos Estados Americanos. **Convenção Americana sobre Direitos Humanos** (Pacto de San José da Costa Rica). San José, 1969.

QUADRADO, Jaqueline Carvalho; FERREIRA, Ewerton da Silva. Ódio e intolerância nas redes sociais digitais. **Revista Katálysis, Florianópolis**, v. 23, n. 3, p. 419-428, 2020.

ROBERTS, Sarah T. **Behind the Screen**: content moderation in the shadows of social media. New Haven: Yale University Press, 2019. Santa Clara Principles. **Santa Clara Principles on Transparency and Accountability in Content Moderation**. 2021.

SANTOS, Gedielson Gabriel dos; BUENO, Mariza Schuster. Direito à liberdade de expressão ou discurso de ódio nas mídias sociais. **Revista Academia de Direito**, Canoinhas, v. 6, p. 2682-2703, 2024.

SARMENTO, Daniel. A liberdade de expressão e o problema do "hate speech". **Revista de Direito do Estado,** Rio de Janeiro, v. 11, p. 169-198, 2016.

SILVA, Virgílio Afonso da. O proporcional e o razoável. **Revista dos Tribunais,** São Paulo, v. 798, p. 23-50, 2002.

SOUZA, Carlos Affonso; LEMOS, Ronaldo (coords.). **Marco Civil da Internet: construção e aplicação.** Juiz de Fora: Editar Editora Associada Ltda., 2016.

Supremo Tribunal Federal. **STF define parâmetros para responsabilização de plataformas por conteúdos de terceiros**. Brasília, DF: Supremo Tribunal Federal, 26 jun. 2025.

TABOGA, Ana Laura Vilamaior; SANTOS JUNIOR, Randolfo. Influência de redes sociais na saúde mental e autoimagem de adolescentes. **Educação, Ciência e Cultura**, Vila Nova de Gaia, v. 25, n. 1, p. 20-30, 2021.

Tribunal Superior Eleitoral. **Programa Permanente de Enfrentamento à Desinformação.** Brasília, DF: Tribunal Superior Eleitoral, 2024-2025.

TRINDADE, Luiz Valério de Souza. **Discurso de ódio nas redes sociais**. Rio de Janeiro: Jandaíra, 2022.

UNESCO. **Diretrizes para a governança das plataformas digitais**. Paris: UNESCO, 2023.

União Europeia. Regulamento (UE) 2022/2065 do Parlamento Europeu e do Conselho, de 19 de outubro de 2022, relativo a um mercado único de serviços digitais e que altera a Diretiva 2000/31/CE (Digital Services Act). **Jornal Oficial da União Europeia**, L 277, 27 out. 2022.

WACHS, Sebastian; WRIGHT, Michelle F.; GÁMEZ-GUADIX, Manuel. From hate speech to HateLess: the effectiveness of a prevention program on adolescents' online hate speech involvement. **Computers in Human Behavior,** v. 157, art. 108250, 2024.