

LIBERDADE DE EXPRESSÃO VERSUS DISCURSO DE ÓDIO NAS MÍDIAS DIGITAIS: dinâmicas sociotécnicas e implicações político-comunicacionais¹

FREEDOM OF EXPRESSION VERSUS HATE SPEECH IN DIGITAL MEDIA: sociotechnical dynamics and political- communicational implications

Érica Anita Baptista²
Bruna Presmic³
Vanessa Marques Castro⁴

Resumo: O ambiente digital transformou as dinâmicas comunicacionais, gerando tensão entre a defesa da liberdade de expressão e a necessidade de contenção do discurso de ódio. Esse artigo busca identificar a ambivalência com foco nas implicações para a opinião pública. Neste sentido, analisa como elementos característicos da esfera digital como o anonimato, a velocidade da informação e os algoritmos, favorecem a difusão de discursos intolerantes. Recorre a teorias da comunicação, como a *Espiral do Silêncio* (Noelle-Neumann, 2019) e os *Públicos de Tarde* (Tarde, 2005), para compreender os processos de polarização e silenciamento simbólico nas redes. A pesquisa visa contribuir para o debate contemporâneo sobre a regulação da comunicação digital, buscando equilibrar a proteção da liberdade expressiva e o enfrentamento a manifestações discriminatórias.

Palavras-Chave: Liberdade de expressão 1. Discurso de ódio 2. Polarização política 3.

Abstract: The digital environment has transformed communicational dynamics, generating tension between the defense of freedom of expression and the need to contain hate speech. This article seeks to identify this ambivalence with a focus on its implications for public opinion. In this sense, it analyzes how characteristic elements of the digital sphere—such as anonymity, information speed, and algorithms—favor the diffusion of intolerant discourse. It draws on communication theories, such as the *Spiral of Silence* (Noelle-Neumann, 2019) and *Tarde's Publics* (Tarde, 2005), to understand the processes of polarization and symbolic silencing in networks. The research aims to

¹Trabalho apresentado ao Grupo de Trabalho Materialidades Digitais e Práticas Comunicacionais. 35º Encontro Anual da Compós, Universidade Federal do Rio Grande do Norte (UFRN). Natal - RN. 9 a 12 de junho de 2026.

² Professora do Programa de Pós-graduação em Comunicação do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). Instituto Nacional de Ciência e Tecnologia em Disputas e Soberanias Informacionais (INCT-DSI). Doutora em Ciência Política pela Universidade Federal de Minas Gerais (UFMG). E-mail: anitaerica@gmail.com.

³Pontifícia Universidade Católica de Minas Gerais (PUC Minas). Mestre em Comunicação Digital pelo IDP e doutoranda em Ciência da Comunicação pela PUC Minas. E-mail: presmic@gmail.com

⁴Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). Mestre em Indústrias Culturais e Comunicação pela Universidade Politécnica de Valência, na Espanha, e mestre em Comunicação Digital pelo IDP. E-mail: castromarquesv@gmail.com.

contribute to contemporary debate on the regulation of digital communication, seeking to balance the protection of expressive freedom and the confrontation of discriminatory manifestations.

Keywords: *Freedom of expression 1. Hate speech 2. Political polarization 3.*

1. Introdução

O ambiente digital transformou profundamente a comunicação contemporânea e os processos de formação da opinião pública. Ao mesmo tempo em que as plataformas digitais ampliaram o acesso à informação e possibilitaram a pluralização das vozes no espaço público, criaram também um cenário propício à proliferação de conteúdos discriminatórios e à intensificação da polarização política. Essa ambivalência manifesta-se de modo particularmente agudo nas democracias em que a radicalização discursiva tem sido mobilizada como estratégia de disputa política.

A tensão entre a liberdade de expressão, direito fundamental nas sociedades democráticas, e o discurso de ódio, expressão que atenta contra a dignidade humana e a coesão social, exige uma reflexão complexa e contextualizada. A liberdade de expressão, embora essencial, não é um direito absoluto: deve ser exercida em equilíbrio com outros princípios constitucionais, como a igualdade, a não discriminação e a proteção de grupos vulneráveis. Como observa Brugger (2007), manifestações que incitam à violência ou ao preconceito extrapolam os limites da liberdade individual e comprometem direitos coletivos fundamentais.

Nas plataformas digitais, características como o anonimato, a instantaneidade e o alcance global potencializam tais tensões. Suler (2004), ao introduzir o conceito de “efeito de desinibição online”, evidencia como fatores como a invisibilidade, a assincronia e a diluição da autoridade favorecem a emergência de comportamentos extremos. Essa desinibição, por sua vez, é amplificada por sistemas algorítmicos voltados à maximização do engajamento emocional, os quais priorizam conteúdos polarizadores e contribuem para a formação de enclaves ideológicos.

Parte-se do argumento de que o discurso de ódio, nas plataformas digitais, não emerge apenas como expressão individual de intolerância, mas como efeito estrutural das dinâmicas sociotécnicas orientadas pelo engajamento e pela maximização da visibilidade. Nesse contexto, a própria arquitetura comunicacional das plataformas tende a incentivar a intensificação da hostilidade discursiva e a radicalização de posições, tensionando os marcos normativos

tradicionais da liberdade de expressão. O problema central investigado consiste em compreender de que modo essas estruturas mediadoras conciliam, ou deixam de conciliar, a proteção à liberdade expressiva com a contenção de manifestações de ódio, sem comprometer os fundamentos democráticos que sustentam a esfera pública.

O objetivo geral deste estudo é analisar os efeitos das dinâmicas sociotécnicas das plataformas digitais sobre os mecanismos de formação da opinião pública e os limites da liberdade de expressão no contexto contemporâneo. Especificamente, busca-se: identificar os elementos estruturais das plataformas digitais que favorecem a disseminação do discurso de ódio; compreender como tais dinâmicas impactam os processos deliberativos e as condições do debate político; e, por fim, examinar propostas contemporâneas de enfrentamento ao discurso de ódio à luz das tensões entre regulação comunicacional e liberdade de expressão.

Para tanto, recorre-se a uma abordagem teórico-conceitual interdisciplinar, ancorada em aportes da teoria da comunicação, da ciência política, do direito e dos estudos sobre plataformas digitais. As teorias da espiral do silêncio de Elisabeth Noelle-Neumann (2019) e dos públicos de Tarde (2005) oferecem chaves interpretativas para compreender os mecanismos de silenciamento simbólico e a formação de coletivos ideológicos extremos. Enquanto a espiral do silêncio explica o silenciamento de opiniões percebidas como dissidentes, as contribuições de Tarde ajudam a compreender como a estrutura técnica das redes favorece a constituição de “públicos de ódio” coesos e persistentes. Ao articular essas perspectivas teóricas com a análise das dinâmicas algorítmicas e das estratégias comunicacionais contemporâneas, este trabalho busca contribuir para o debate sobre a regulação democrática da comunicação digital, com vistas a promover um ambiente informacional mais plural, inclusivo e respeitoso das diferenças.

Nesse contexto, o discurso de ódio é compreendido neste artigo não apenas como manifestação isolada de intolerância ou violação normativa, mas como componente estrutural de ecossistemas informacionais marcados pela polarização, pela moralização do debate público e pela circulação estratégica de conteúdos enganosos. Ao operar por meio da desumanização, da construção de inimigos e do silenciamento simbólico de grupos sociais, o discurso de ódio estabelece condições cognitivas e afetivas favoráveis à desinformação, contribuindo para a degradação da esfera pública e para o enfraquecimento dos processos deliberativos nas democracias contemporâneas.

Este artigo se organiza em quatro seções. A primeira delas traz as bases metodológicas que foram desenhadas a partir da complexidade do objeto de estudo. A seção seguinte apresenta e discute sobre os mecanismos digitais de ampliação do discurso de ódio, a saber: a percepção do anonimato; a velocidade e o alcance da disseminação de conteúdos nas plataformas digitais; os algoritmos de personalização; "dinâmica de grupo online"; e a economia da atenção. Na terceira seção, analisamos e discutimos algumas propostas que buscam alcançar um ponto de equilíbrio entre o combate às manifestações de intolerância e ódio e a garantia das liberdade de expressão. Por fim, na quarta e última seção, retomamos os principais pontos abordados neste trabalho, discutimos sobre os limites da pesquisa e indicamos as fases seguintes da investigação.

2. Metodologia

Este estudo adota uma abordagem teórico-conceitual, de natureza exploratória e crítica, voltada à análise das inter-relações entre liberdade de expressão, discurso de ódio e as dinâmicas sociotécnicas das plataformas digitais. A escolha metodológica fundamenta-se na natureza complexa e multifacetada do objeto de estudo, cuja compreensão exige articulação entre diferentes campos do conhecimento e aprofundamento conceitual. A análise desenvolve-se a partir de uma revisão crítica e interpretativa da literatura especializada, orientada pela construção conceitual do objeto.

A pesquisa se ancora em uma perspectiva interdisciplinar, integrando aportes teóricos oriundos da Comunicação, do Direito, da Sociologia, da Ciência Política e da Psicologia Social. Cada campo contribui com ferramentas analíticas específicas: a Comunicação fornece os referenciais para compreender os processos de visibilidade, formação da opinião pública e circulação simbólica; o Direito aporta o arcabouço normativo que regula as tensões entre direitos fundamentais; a Sociologia digital permite compreender as práticas e interações sociais mediados por tecnologia; a Ciência Política oferece perspectivas sobre deliberação democrática e polarização; e a Psicologia Social contribui com a análise dos mecanismos cognitivos e afetivos que impulsionam a radicalização discursiva.

Metodologicamente, o estudo está estruturado em uma revisão crítica da literatura especializada, voltada à identificação, seleção e análise interpretativa de obras clássicas e contemporâneas que tratam dos temas centrais em debate. Seguiu-se as diretrizes gerais de Creswell (2010), com foco em: (i) delimitação de palavras-chave; (ii) levantamento em bases

de dados acadêmicas reconhecidas; (iii) seleção criteriosa das fontes com base em relevância temática e rigor acadêmico; (iv) leitura e análise crítica do conteúdo; e (v) construção de uma síntese interpretativa articulada aos objetivos do estudo.

Foram consideradas estudos clássicos e contemporâneos que tratam dos temas centrais em debate, buscando articular diferentes tradições teóricas e perspectivas analíticas. Como forma de sistematizar e padronizar os dados, foram selecionadas as publicações em periódicos de alto impacto, livros de referência teórica consolidada e relatórios produzidos por organismos internacionais e centros de pesquisa. O *corpus* inclui autores como Elisabeth Noelle-Neumann (2019), Gabriel Tarde (2005), John Suler (2004), Cass Sunstein (2017), Manuel Castells (2015), entre outros, além de referências brasileiras e latino-americanas que abordam a especificidade dos processos comunicacionais na região. Também foram considerados os trabalhos empíricos realizados, especialmente sobre o contexto brasileiro, nos últimos dez anos. Esse recorte temporal considera, sobretudo, as mudanças no contexto político a partir do impeachment de Dilma Rousseff, período em que é possível identificar um aumento expressivo nessa agenda.

Reconhecem-se, entretanto, as limitações da abordagem adotada. A ausência de coleta de dados primários restringe a possibilidade de verificação empírica das hipóteses formuladas, embora a densidade conceitual construída sirva como base para futuros estudos empíricos. O caráter dinâmico e mutável do ambiente digital, com mudanças constantes nas plataformas, algoritmos e regulações, impõe um desafio à fixação analítica do objeto. Ademais, apesar do esforço por incluir referências regionais, o escopo da bibliografia consultada permanece majoritariamente centrado em produções norte-americanas e europeias, o que poderá ser corrigido em investigações futuras. Ainda assim, acredita-se que a abordagem teórica interdisciplinar adotada oferece um panorama robusto e fundamentado sobre as implicações político-comunicacionais das plataformas digitais, contribuindo para o avanço do debate acadêmico e a formulação de estratégias normativas de enfrentamento ao discurso de ódio.

3. Mecanismos digitais de ampliação do discurso de ódio

O ambiente digital apresenta características estruturais que potencializam a disseminação e o impacto do discurso de ódio, criando dinâmicas significativamente distintas daquelas observadas nos meios de comunicação tradicionais. Compreender esses mecanismos

de ampliação é fundamental para desenvolver estratégias eficazes de enfrentamento do problema. Neste estudo, propõe-se a análise de cinco mecanismos principais que contribuem para a amplificação do discurso de ódio no ambiente digital.

O primeiro mecanismo relevante é o anonimato, ou a percepção de anonimato, proporcionado pelas plataformas digitais. A possibilidade de expressar opiniões sem revelar a identidade real reduz significativamente as barreiras sociais que normalmente inibem comportamentos agressivos ou discriminatórios (Citron, 2014). Esse fenômeno está intimamente relacionado ao que Suler (2004) denominou "efeito de desinibição online", caracterizado pela tendência dos indivíduos a comportarem-se de maneira mais extrema e desinibida no ambiente virtual.

O segundo mecanismo significativo é a velocidade e o alcance da disseminação de conteúdos nas plataformas digitais. A arquitetura comunicacional das plataformas digitais viabiliza a circulação instantânea de conteúdos em escala global, permitindo que mensagens, inclusive aquelas marcadas por hostilidade discursiva, alcancem rapidamente audiências ampliadas e heterogêneas (Castells, 2015). Essa característica é potencializada pela cultura da convergência, na qual conteúdos fluem facilmente entre diferentes plataformas e formatos, amplificando seu alcance e impacto (Jenkins, 2013).

O terceiro mecanismo de ampliação refere-se aos algoritmos de personalização utilizados pelas plataformas digitais, projetados para maximizar o engajamento dos usuários. Esses algoritmos tendem a priorizar conteúdos que provocam reações emocionais intensas e a mostrar aos usuários informações que reforçam suas crenças e preferências existentes, formando as chamadas bolhas de filtro (Pariser, 2011). Esse fenômeno pode levar à radicalização, à medida que os usuários são repetidamente expostos a discursos extremos que ressoam com suas próprias opiniões, sem contato significativo com perspectivas alternativas.

O quarto mecanismo relevante é o que podemos denominar "dinâmica de grupo online". As plataformas digitais facilitam a formação de comunidades baseadas em interesses ou visões compartilhadas, nas quais indivíduos com opiniões semelhantes podem reforçar mutuamente suas crenças e comportamentos (Crockett e Brady et al., 2017). Esse fenômeno está relacionado à polarização de grupo.

Por fim, um quinto mecanismo de ampliação refere-se à economia da atenção (Byung-Chul Han, 2017). que caracteriza o modelo de negócios das plataformas digitais. Os conteúdos

que provocam reações emocionais intensas, incluindo indignação, raiva e ódio, tendem a gerar maior engajamento e, conseqüentemente, maior valor econômico. Esse modelo cria incentivos estruturais para a proliferação de conteúdos polarizadores e sensacionalistas, incluindo discursos de ódio. Como demonstra estudo conduzido por Crockett (2017), expressões de indignação moral nas redes sociais são recompensadas com maior engajamento (curtidas, compartilhamentos, comentários), criando um ciclo de reforço que favorece a escalada da hostilidade no discurso público.

A combinação desses cinco mecanismos, anonimato, velocidade e alcance da disseminação, algoritmos de personalização, dinâmicas de grupo e economia da atenção, cria um ambiente particularmente propício para a proliferação do discurso de ódio nas plataformas digitais. Compreender essas dinâmicas é essencial para desenvolver estratégias eficazes de mitigação do problema, que necessariamente devem abordar tanto aspectos técnicos quanto sociais e econômicos. Nesse sentido, o discurso de ódio não deve ser compreendido como uma disfunção externa às plataformas digitais, mas como um fenômeno compatível com modelos de negócio orientados pela captura e monetização da atenção (Vaidhyanathan, 2018).

Sob esse olhar, o discurso de ódio não pode ser dissociado das dinâmicas desinformativas que atravessam o ambiente digital. Frequentemente, manifestações odiosas são sustentadas por informações falsas, distorcidas ou descontextualizadas, que operam como gatilhos emocionais e reforçam enquadramentos morais binários. Assim, o discurso de ódio funciona como uma infraestrutura simbólica da desinformação, ao reduzir a complexidade dos fenômenos sociais, inviabilizar o contraditório e produzir ambientes cognitivos hostis à verificação e à argumentação racional. Mais do que fatores isolados, esses mecanismos operam de forma combinada, produzindo um ambiente comunicacional no qual a visibilidade pública passa a depender da intensidade afetiva das interações, criando incentivos estruturais à radicalização discursiva.

Bolhas ideológicas e radicalização discursiva

Os algoritmos de personalização representam um dos aspectos mais controversos e impactantes do ecossistema informacional contemporâneo. Projetados para filtrar e priorizar conteúdos com base nas preferências e comportamentos prévios dos usuários, esses algoritmos

têm consequências significativas para a formação da opinião pública e para a disseminação do discurso de ódio.

As bolhas ideológicas têm implicações diretas nesse processo. Como observa Sunstein (2017), quando indivíduos interagem predominantemente com outros que compartilham suas visões, três mecanismos principais contribuem para a radicalização: o pool limitado de argumentos (exposição predominante a argumentos que apoiam a visão inicial), a comparação social (desejo de ser percebido positivamente pelo grupo) e a confiança reforçada (maior segurança nas próprias opiniões quando apoiadas por outros).

Esse processo pode levar à normalização do discurso de ódio dentro de comunidades específicas. Como demonstra estudo conduzido por Magdalena Wojcieszak (2010), a participação em grupos online ideologicamente homogêneos está associada a maior extremismo e menor tolerância a perspectivas divergentes. Quando manifestações discriminatórias são frequentes e recebem aprovação dentro de uma comunidade, elas tendem a ser percebidas como aceitáveis ou mesmo desejáveis.

Nessa perspectiva, o discurso de ódio não pode ser dissociado das dinâmicas desinformativas que atravessam o ambiente digital. Frequentemente, manifestações odiosas são sustentadas por informações falsas, distorcidas ou descontextualizadas, que operam como gatilhos emocionais e reforçam enquadramentos morais binários. Assim, o discurso de ódio funciona como uma infraestrutura simbólica da desinformação, ao reduzir a complexidade dos fenômenos sociais, inviabilizar o contraditório e produzir ambientes cognitivos hostis à verificação e à argumentação racional.

Arquitetura algorítmica e incentivos estruturais

É importante ressaltar que os algoritmos de personalização não são neutros, mas refletem escolhas de design e prioridades comerciais. Nesse sentido, o modelo de negócios das plataformas digitais, baseado na captura e monetização da atenção dos usuários, reforça a formação de bolhas ideológicas e, consequentemente, a disseminação do discurso de ódio. Vaidhyanathan (2018), ao analisar o Facebook, argumenta que a lógica da plataforma é desenhada para maximizar reações emocionais intensas, mesmo quando essas reações são mobilizadas por conteúdos odiosos.

Diante desses desafios, pesquisadores como Bozdag e van den Hoven (2015) propõem o desenvolvimento de "algoritmos de diversidade", que exporiam os usuários a diferentes perspectivas. De modo semelhante, Helberger et al. (2018) sugerem que plataformas incorporem a “diversidade por design” como um princípio de engenharia informacional.

Polarização política e campanhas digitais

Como observam Jennifer Stromer-Galley (2014) e Kreiss (2016), as campanhas políticas contemporâneas utilizam intensivamente dados e estratégias de microsegmentação para direcionar mensagens específicas a diferentes públicos. Essa personalização pode ser eficaz para engajamento, mas também contribui para a fragmentação do debate público.

Rossini et al. (2016) demonstram que, em contextos de alta polarização política, as campanhas digitais são amplamente utilizadas pelos cidadãos para debater, confrontar e disseminar visões polarizadas. Embora as mensagens oficiais visem à construção da imagem dos candidatos, a apropriação por parte dos eleitores frequentemente acirra os conflitos discursivos. Nesse cenário, as informações recebidas se moldam às preferências dos cidadãos, que recusam contato ou descredibilizam quaisquer conteúdos que, porventura, sejam distantes de seu polo afetivo.

A polarização política é um fenômeno complexo e que pode se intensificar de acordo com a situação política do país. No Brasil, as campanhas eleitorais presidenciais colocam em evidência a polarização, sobretudo, a partir de 2018. Apesar de ser um reflexo do contexto, a polarização não necessariamente tem relação direta com o interesse e com a confiança na política, ou nas instituições políticas democráticas. Mas pode se intensificar diante dessas situações. A literatura nos mostra pelo menos duas variantes que podem ou não estar diretamente relacionadas: a polarização ideológica e a polarização afetiva. A polarização ideológica seria o aumento da distância entre as posições e crenças de diferentes grupos ou indivíduos sobre questões políticas, sociais ou culturais (Dias e Germano et al., 2024; Fucks, Marques, 2022). Quando ocorre um insulamento dessas diferenças entre grupos sociais, associamos à polarização afetiva. Os grupos enxergam de maneira enviesada, segundo a sua lente social, tanto os atributos dos membros de seu grupo quanto dos grupos oponentes (Nunes e Traumann, 2023; Marques, 2025). Assim, os dados da realidade objetiva são percebidos e analisados a partir de um filtro emocional (Dias e Germano, et al. 2024; Nunes e Traumann,

2023). No ambiente das plataformas digitais, essa distinção tende a se recompor, uma vez que mecanismos algorítmicos convertem divergências cognitivas em antagonismos afetivos ao amplificar seletivamente conteúdos moralmente carregados e emocionalmente intensos.

Ambivalências das plataformas digitais: riscos e oportunidades

As plataformas digitais, por suas características técnicas, impõem novos desafios ao controle da mensagem política. Em cenários competitivos, como lembra Iyengar et al. (2019), a polarização afetiva torna-se mais evidente: eleitores passam a ver adversários não como oponentes legítimos, mas como inimigos morais, intensificando a intolerância e a hostilidade (Marques, 2025). Por outro ângulo, é possível entender a polarização, nomeadamente a afetiva, associada a mecanismos psicossociais de identidade social.

A emergência da polarização do tipo afetiva na sociedade das mídias sociais pode ser compreendida por meio de modelos que ligam a tecnologia da informação a mecanismos psicossociais de identidade social. Sob esse recorte, Törnberg et al. (2021) sugerem que as mídias digitais, ao deslocarem a interação de um contexto primariamente local para um ambiente menos restrito no espaço, permitem que indivíduos se agrupem com outros semelhantes⁵. Este processo, ainda que seja um limitador de visões divergentes e do diálogo, fortalece as identidades sociais salientes e reduz o terreno comum com aqueles que pensam de forma diferente. Esse comportamento pode reduzir desgastes e conflitos, porém, também favorece o declínio do diálogo e das trocas argumentativas.

No entanto, é preciso reconhecer que as redes também oferecem oportunidades para a participação democrática. Gil de Zúñiga et al. (2012) e Bimber et al. (2015) destacam que o uso político das mídias sociais pode fortalecer formas de engajamento cívico, mobilização coletiva e deliberação pública. A capacidade de disseminar informações rapidamente, organizar movimentos sociais e dar voz a grupos marginalizados representa um ganho significativo para a democracia. A tensão reside, portanto, na gestão dessas ambivalências, buscando maximizar os benefícios democráticos enquanto se mitigam os riscos associados à disseminação de conteúdos prejudiciais e à fragmentação social.

⁵ A importância do contato com opiniões divergentes e do diálogo é destacada em diversos trabalhos. Aqui citamos os trabalhos de Lev-on e Manin (2009), que, já em 2009, discutiam a formação das “bolhas” e seus impactos na sociedade.

4. Análise e discussão

Diante dos desafios apresentados pelo discurso de ódio nas plataformas digitais, diversos pesquisadores, organizações da sociedade civil e formuladores de políticas têm proposto abordagens inovadoras que buscam equilibrar a proteção contra manifestações discriminatórias com a preservação da liberdade de expressão. Nesta seção, examinamos algumas dessas propostas e perspectivas.

Uma abordagem promissora é o que Kate Klonick (2020) denomina "governança de plataformas", que enfatiza a necessidade de estruturas de governança mais transparentes, participativas e *accountability* para as decisões de moderação de conteúdo. A autora argumenta que as plataformas digitais funcionam cada vez mais como "governos privados", exercendo poder significativo sobre o discurso público, e que esse poder deve ser sujeito a princípios de legitimidade democrática.

O *Oversight Board* da Meta, estabelecido em 2020, representa uma tentativa de implementar alguns desses princípios. Composto por especialistas independentes em áreas como liberdade de expressão, direitos humanos e tecnologia, o conselho tem autoridade para revisar e potencialmente reverter decisões de moderação de conteúdo da plataforma. Como observa Evelyn Douek (2021), embora o *Oversight Board* represente um avanço em termos de transparência e *due process*, seu escopo limitado e sua dependência financeira da Meta levantam questões sobre sua eficácia como mecanismo de *accountability*.

Outra proposta significativa é o que Gorwa (2019) denomina "regulação baseada em auditoria" (*audit-based regulation*). Nessa abordagem, reguladores estabelecem padrões de desempenho para plataformas em áreas como proteção contra discurso de ódio, mas delegam a verificação de conformidade a auditores independentes. Isso permitiria combinar a expertise técnica necessária para avaliar sistemas complexos com a legitimidade democrática da supervisão regulatória.

Complementarmente, Pasquale (2015) propõe maior "transparência algorítmica", argumentando que os algoritmos que moldam o fluxo de informações nas plataformas digitais deveriam estar sujeitos a escrutínio público e regulatório. O autor sugere que reguladores deveriam ter acesso aos códigos-fonte e conjuntos de dados utilizados pelas plataformas,

permitindo avaliar se esses sistemas inadvertidamente amplificam discursos de ódio ou outros conteúdos prejudiciais.

Uma perspectiva distinta é oferecida por Balkin (2018), que propõe o conceito de "constitucionalismo de plataforma" (*platform constitutionalism*). O autor argumenta que, assim como Estados desenvolveram constituições para limitar e legitimar seu poder, as plataformas digitais deveriam adotar princípios constitucionais que estabeleçam direitos dos usuários, limites ao poder das plataformas e mecanismos de participação na governança. Esses princípios poderiam ser implementados por meio de uma combinação de autorregulação, pressão de mercado e intervenção estatal.

No contexto brasileiro, Belli e Venturini (2020) propõem uma abordagem de "regulação responsiva", que combinaria incentivos para autorregulação efetiva com a possibilidade de intervenção regulatória mais rigorosa quando necessário. Os autores argumentam que essa abordagem seria particularmente adequada ao contexto brasileiro, caracterizado por alta desigualdade digital e instituições democráticas ainda em consolidação.

Uma perspectiva complementar é oferecida por iniciativas focadas no letramento digital e na resiliência social. Como observa Wardle (2019), estratégias puramente regulatórias ou tecnológicas são insuficientes para abordar problemas complexos como o discurso de ódio online. A autora propõe uma abordagem ecossistêmica que inclui educação midiática, fortalecimento do jornalismo de qualidade e desenvolvimento de normas sociais que desencorajem a disseminação de conteúdos prejudiciais.

Nessa linha, o projeto "*Counterspeech*" da *Dangerous Speech Project*, liderado por Susan Benesch (2014), enfoca o desenvolvimento de estratégias para contestar efetivamente discursos de ódio sem recorrer à censura. A iniciativa baseia-se em evidências de que intervenções de "contradiscurso" (*counterspeech*), respostas não violentas que desafiam estereótipos, corrigem desinformação e promovem empatia, podem ser eficazes em reduzir o impacto de manifestações discriminatórias.

Por fim, uma perspectiva emergente enfoca o redesenho das próprias plataformas digitais para minimizar incentivos à polarização e ao discurso de ódio. Como argumenta Zuckerman (2021), o modelo atual de redes sociais, otimizado para maximizar engajamento e tempo de permanência, cria incentivos estruturais para conteúdos polarizadores e sensacionalistas. O autor propõe o desenvolvimento de "redes sociais públicas" (*public social*

media), projetadas para promover deliberação democrática e coesão social em vez de maximizar lucros publicitários. Essas diversas propostas e perspectivas refletem a complexidade do desafio representado pelo discurso de ódio nas plataformas digitais. Não existe uma solução única e definitiva para esse problema (Suzor, 2019).

Nesse sentido, as iniciativas de governança, transparência algorítmica e regulação das plataformas digitais devem ser compreendidas não apenas como mecanismos de contenção do discurso de ódio, mas como estratégias voltadas à proteção da integridade da informação. Ao enfrentar incentivos estruturais que favorecem a viralização de conteúdos extremos e enganosos, tais políticas contribuem para a preservação das condições mínimas de um debate público informado, plural e democraticamente orientado.

5. Considerações finais

Esse cenário tende a se intensificar com o avanço da inteligência artificial nos processos de produção, recomendação e moderação de conteúdo. A automação crescente da comunicação digital amplia a escala, a velocidade e a sofisticação da circulação de discursos de ódio e de conteúdos enganosos, ao mesmo tempo em que dificulta a identificação de responsabilidades e a aplicação de mecanismos tradicionais de controle. A capacidade de sistemas de IA produzirem enunciados altamente verossímeis e performativos introduz novos desafios à confiança pública e à integridade do ecossistema informacional, demandando atenção crescente da pesquisa acadêmica e das instâncias regulatórias.

O desafio, portanto, não é rejeitar as plataformas como arenas políticas, mas sim desenvolver estratégias que ampliem seu potencial democrático e limitem seus efeitos nocivos. Isso passa por ações em múltiplas frentes: desenho de plataformas mais responsáveis, regulação proporcional, educação midiática e compromisso ético dos atores políticos. A tensão entre liberdade de expressão e proteção contra discursos prejudiciais continuará a representar um dos desafios mais significativos. Navegar esse terreno complexo requer não apenas análises rigorosas e políticas bem fundamentadas, mas também um compromisso coletivo com valores democráticos fundamentais.

A análise desenvolvida evidencia que o discurso de ódio nas plataformas digitais não representa meramente uma extensão de manifestações discriminatórias preexistentes, mas adquire características específicas no ambiente online. Evidenciando assim, a urgência de

compreender e combater essa prática diante das implicações e isso requer uma abordagem interdisciplinares e regulatórias eficazes para mitigar os impactos negativos.

Podemos concluir que o enfrentamento ao discurso de ódio nas plataformas digitais requer uma abordagem multidimensional, que combine intervenções regulatórias, tecnológicas, educacionais e sociais. O objetivo central não deve ser a eliminação completa do discurso de ódio, um objetivo provavelmente inatingível, e sim a criação de um ecossistema digital que minimize seus danos e maximize as oportunidades para expressão construtiva e deliberação democrática. Os próximos passos desta pesquisa envolvem a análise de discursos nas redes sociais online em diferentes momentos políticos, tendo em vista a intensidade da polarização, e a posterior comparação dos dados, tendo em vista o tensionamento entre o discurso de ódio e a liberdade de expressão, e os mecanismos digitais que oportunizam o discurso de ódio, como discutido neste artigo.

Acreditamos, ainda, que futuras investigações devem aprofundar-se nas estratégias de enfrentamento e na educação digital para promover um ambiente online mais inclusivo e respeitoso. Embora a eliminação das agressões a partir do discurso de ódio não pareça ser um horizonte possível, ações interdisciplinares de educação e combate à naturalização desse tipo de comportamento precisam estar nas agendas de pesquisa e se concretizar em políticas efetivas.

Em síntese, o discurso de ódio nas plataformas digitais deve ser compreendido como fenômeno sociotécnico indissociável das dinâmicas contemporâneas de desinformação. Sua articulação com regimes algorítmicos de visibilidade, economia da atenção e polarização afetiva amplia os riscos à deliberação democrática e à formação de opiniões informadas. O enfrentamento desse quadro exige abordagens integradas, que articulem regulação, desenho responsável das plataformas, educação midiática e compromisso ético dos atores políticos, sob pena de normalização de práticas que corroem a esfera pública e fragilizam as democracias. As perspectivas analisadas indicam que o enfrentamento do discurso de ódio exige compreender as plataformas não apenas como meios de circulação discursiva, mas também como ambientes regulatórios que reorganizam as condições de produção da visibilidade pública.

Referências

- BALKIN, Jack M. Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. **UC Davis Law Review**, v. 51, p. 1149-1210, 2018.
- BELLI, L.; VENTURINI, J. Private ordering and the rise of terms of service as cyber-regulation. **Internet Policy Review**, v. 5, n. 4, 2016.
- BENESCH, Susan. **Countering Dangerous Speech: New Ideas for Genocide Prevention**. Washington, DC: United States Holocaust Memorial Museum, 2014.
- BIMBER, B.; FLANAGIN, A. J.; STOHL, C. **Collective Action in Organizations: Interaction and Engagement in an Era of Technological Change**. Cambridge: Cambridge University Press, 2012.
- BOZDAG, E.; VAN DEN HOVEN, J. Breaking the filter bubble: democracy and design. **Ethics and Information Technology**, v. 17, n. 4, p. 249-265, 2015.
- BRADY, W.; WILLS, A.; JOST, J. T.; TUCKER, J. A. VAN BAVEL, J. J. et al. Emotion shapes the diffusion of moralized content in social networks. **Proceedings of the National Academy of Sciences**, v. 114, n. 28, p. 7313-7318, 2017.
- BRUGGER, Winfried. Proibição ou proteção do discurso do ódio? Algumas observações sobre o direito alemão e o americano. **Direito Público**, v. 4, n. 15, p. 117-136, 2007.
- CASTELLS, Manuel. **O Poder da Comunicação**. São Paulo/Rio de Janeiro: Paz e Terra, 2015.
- CITRON, Danielle K. **Hate Crimes in Cyberspace**. Cambridge: Harvard University Press, 2014.
- CRESWELL, John W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. Porto Alegre: Artmed, 2010.
- DIAS, M. R.; GERMANO, L. B.; SILVA, M. K.; MACHADO, S. R. Polarização Afetiva e Polarização Ideológica: determinantes da participação política e da abstenção eleitoral. **48º Encontro Anual da Anpocs**, Campinas (São Paulo), 16 a 25 de outubro, 2024.
- DOUEK, Evelyn. The Limits of International Law in Content Moderation. **UC Irvine Journal of International, Transnational, and Comparative Law**, v. 6, p. 37-78, 2021.
- FUKS, M.; MARQUES, P. H. Polarização e contexto: medindo e explicando a polarização política no Brasil. **OPINIÃO PÚBLICA**, Campinas, vol. 28, nº 3, p. 560-593, set.-dez., 2022.
- GIL DE ZÚÑIGA, H.; JUNG, N.; VALENZUELA, S. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. **Journal of Computer-Mediated Communication**, v. 17, n. 3, p. 319-336, 2012.
- GORWA, Robert. The platform governance triangle: conceptualising the informal regulation of online content. **Internet Policy Review**, v. 8, n. 2, 2019.
- HAN, Byung-Chul. **Sociedade do cansaço**. Petrópolis: Vozes, 2017.
- HELBERGER, N.; KARPPINEN, K.; D'ACUNTO, L. Exposure diversity as a design principle for recommender systems. **Information, Communication & Society**, v. 21, n. 2, p. 191-207, 2018.

IYENGAR, S.; LELKES, Y.; LEVENDUSKY, M.; MALHOTRA, N.; WESTWOOD, S. J. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, v. 22, p. 129-146, 2019.

JENKINS, H.; FORD, S.; GREEN, J. **Cultura da Conexão: criando valor e significado por meio da mídia propagável**. São Paulo: Aleph, 2014.

KLONICK, Kate. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *Yale Law Journal*, v. 129, p. 2418-2499, 2020.

KREISS, Daniel. **Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy**. New York: Oxford University Press, 2016.

LEV-ON, Azi; MANIN, Bernard. Happy accidents: Deliberation and online exposure to opposing views. In: Todd Davies and Seeta Gangadharan (eds). **Online Deliberation: Design, Research and Practice**. Chicago: Center for the Study of Language and information, 2009. Pp. 105-122.

MARQUES, V. **Neopopulismo e valores conservadores: a gramática da direita conservadora nas mídias sociais digitais nos Estados Unidos, Brasil e Argentina**. Dissertação (Mestrado em Comunicação Digital) - Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). 150f. Brasília, DF. 2025.

NOELLE-NEUMANN, Elisabeth. **A espiral do silêncio: opinião pública: nossa pele social**. Tradução de Maria Aparecida Baccega. São Paulo: Paulus, 2019.

NUNES, F.; TRAUMANN, T. **Biografia do abismo: Como a polarização divide famílias desafia empresas e compromete o futuro do Brasil**. Rio de Janeiro: Editora HarperCollins, 2023.

PARISER, Eli. **O Filtro Invisível: O que a internet está escondendo de você**. Rio de Janeiro: Zahar, 2012.

PASQUALE, Frank. **The Black Box Society: The Secret Algorithms That Control Money and Information**. Cambridge: Harvard University Press, 2015.

ROSSINI, P. G. C.; BAPTISTA, É. A.; OLIVEIRA, V. V.; SAMPAIO, R. C. O uso do Facebook nas eleições presidenciais brasileiras de 2014: a influência das pesquisas eleitorais nas estratégias das campanhas digitais. *Fronteiras - Estudos Midiáticos*, v. 18, n. 2, p. 145-157, 2016.

SANTAELLA, Lucia. **A pós-verdade é verdadeira ou falsa?** Barueri: Editora Estação das Letras e Cores, 2019.

STROMER-GALLEY, Jennifer. **Presidential Campaigning in the Internet Age**. New York: Oxford University Press, 2014.

SULER, John. The Online Disinhibition Effect. *CyberPsychology & Behavior*, v. 7, n. 3, p. 321-326, 2004.

SUNSTEIN, Cass. R. **#Republic: Divided Democracy in the Age of Social Media**. Princeton: Princeton University Press, 2017.

SUZOR, Nicolas. **Lawless: The Secret Rules That Govern our Digital Lives**. Cambridge: Cambridge University Press, 2019.

TARDE, Gabriel. **A opinião e as massas**. São Paulo: Martins Fontes, 2005.

TÖRNBERG, P.; ANDERSSON, C.; LINDGREN, K.; BANISCH, S. Modeling the emergence of affective polarization in the social media society. *PLoS ONE*, v. 16, n.10, e0258259, 2021.

VAIDHYANATHAN, Siva. **Anti-Social Media: How Facebook Disconnects Us and Undermines Democracy**. New York: Oxford University Press, 2018.

WARDLE, C.; DERAKHSHAN, H. **Information Disorder: Toward an interdisciplinary framework for research and policy making**. Strasbourg: Council of Europe, 2017.

WOJCIESZAK, Magdalena. 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. **New Media & Society**, v. 12, n. 4, p. 637-655, 2010.

ZUCKERMAN, Ethan. **The Case for Digital Public Infrastructure**. Knight First Amendment Institute, Columbia University, 2020.