XXXII CONGRESSO NACIONAL DO CONPEDI SÃO PAULO - SP

INTERNET: DINÂMICAS DA SEGURANÇA PÚBLICA E INTERNACIONAL

DANIELLE JACON AYRES PINTO

GUSTAVO RABAY GUERRA

JOSÉ RENATO GAZIERO CELLA

JÉSSICA FACHIN

I61

Internet: dinâmicas da segurança pública internacional[Recurso eletrônico on-line] organização CONPEDI

Coordenadores: Danielle Jacon Ayres Pinto, Gustavo Rabay Guerra, José Renato Gaziero Cella, Jéssica Fachin – Florianópolis: CONPEDI, 2025.

Inclui bibliografia
ISBN: 978-65-5274-285-8
Modo de acesso: www.conpedi.org.br em publicações
Tema: Os Caminhos Da Internacionalização E O Futuro Do Direito

1. Direito – Estudo e ensino (Pós-graduação) – Encontros Nacionais. 2. Internet. 3. Segurança pública internacional. XXXII Congresso Nacional do CONPEDI São Paulo - SP (4: 2025: Florianópolis, Brasil).

CDU: 34

XXXII CONGRESSO NACIONAL DO CONPEDI SÃO PAULO - SP

INTERNET: DINÂMICAS DA SEGURANÇA PÚBLICA E INTERNACIONAL

Apresentação

No XXII Congresso Nacional do CONPEDI, realizado nos dias 26, 27 e 28 de novembro de 2025, o Grupo de Trabalho - GT “Internet: Dinâmicas da Segurança Pública e Internacional”, que teve lugar na tarde de 28 de novembro de 2025, destacou-se no evento não apenas pela qualidade dos trabalhos apresentados, mas pelos autores dos artigos, que são professores pesquisadores acompanhados de seus alunos pós-graduandos. Foram apresentados artigos objeto de um intenso debate presidido pelos coordenadores.

Esse fato demonstra a inquietude que os temas debatidos despertam na seara jurídica. Cientes desse fato, os programas de pós-graduação em direito empreendem um diálogo que suscita a interdisciplinaridade na pesquisa e se propõe a enfrentar os desafios que as novas tecnologias impõem ao direito. Para apresentar e discutir os trabalhos produzidos sob essa perspectiva.

Os artigos que ora são apresentados ao público têm a finalidade de fomentar a pesquisa e fortalecer o diálogo interdisciplinar em torno do tema “Internet: Dinâmicas da Segurança Pública e Internacional”. Trazem consigo, ainda, a expectativa de contribuir para os avanços do estudo desse tema no âmbito da pós-graduação em direito, apresentando respostas para uma realidade que se mostra em constante transformação.

Os Coordenadores

Prof. Dr. José Renato Gaziero Cella

Prof. Dra. Danielle Jacon Ayres Pinto

Prof. Dr. Gustavo Rabay Guerra

Prof. Dra. Jéssica Fachin

DA MODERAÇÃO DE CONTEÚDO ON-LINE À PREVENÇÃO EDUCATIVA: COMBATE AO DISCURSO DE ÓDIO E DISCURSO PERIGOSO NO INSTAGRAM

FROM ON-LINE CONTENT MODERATION TO EDUCATIONAL PREVENTION: COUNTERING HATE SPEECH AND DANGEROUS SPEECH ON INSTAGRAM

Thayane Brito de Jesus [1]
Felipe Marchese [2]
Maria Amélia Carvalho Campos [3]

Resumo

O artigo investiga se a moderação de conteúdo do Instagram, tomada isoladamente, é capaz de conter a recorrência do discurso de ódio. A pesquisa adota uma metodologia qualitativa, de natureza teórico-normativa, estruturada em revisão bibliográfica e com abordagem dedutiva. Verificou-se, partindo da transformação da esfera pública pela plataformização, que algoritmos e decisões de design funcionam como gatekeepers que estruturam visibilidade, alcance e velocidade da comunicação. Adotou-se, analiticamente, a categoria de discurso perigoso, centrada em efeitos e contexto (orador, audiência, mensagem, meio e ambiente), para graduar risco e orientar respostas mais finas ao combate a discursos potencialmente violentos on-line. A partir da análise das medidas adotados pelo Instagram, o estudo critica a baixa densidade operacional de suas diretrizes contra o discurso de ódio e a tensão entre integridade do ambiente e incentivos de engajamento, defendendo que a melhor solução perpassa pela procedimentalização de garantias de transparência, motivação, recurso e auditoria, bem como pela governança de fluxos com medidas graduais, além de investimentos em educação digital de longo prazo por meio da alfabetização midiática e da literacia algorítmica como estratégia para reduzir a suscetibilidade e a cumulatividade dos danos provocados pelo discurso perigoso nas plataformas digitais. Conclui-se que a moderação é necessária, porém insuficiente de forma isolada, sendo imprescindível um arranjo policêntrico que acople direito, engenharia de plataformas e formação cívica para combater adequadamente o discurso de ódio num contexto de autotransformação da comunicação coletiva.

[1] Advogada e Mestranda em Direito Político e Econômico pela Universidade Presbiteriana Mackenzie (UPM), em associação com a Universidade Federal de Mato Grosso do Sul (UFMS). Técnica de Nível Superior da Universidade Estadual de Mato Grosso do Sul (UEMS). E-mail: thayanebrito@outlook.com

[2] Advogado e Mestrando em Direito Político e Econômico pela Universidade Presbiteriana Mackenzie (UPM), em associação com a Universidade Federal de Mato Grosso do Sul (UFMS). E-mail: felipe.marchese@outlook.com.br.

[3] Advogada e Mestranda em Direito Político e Econômico pela Universidade Presbiteriana Mackenzie (UPM), em associação com a Universidade Federal de Mato Grosso do Sul (UFMS). E-mail: amelia.campos@ufms.br

Palavras-chave: Discurso de ódio, Discurso perigoso, Governamentalidade algorítmica, Moderação de conteúdo, Educação digital

Abstract/Resumen/Résumé

The article aims to evaluate the effectiness of Instagram's content moderation, taken in isolation, is capable of containing the recurrence of hate speech. The research adopts a qualitative methodology, with a theoretical-normative orientation, structured on bibliographic review and a deductive approach. Starting from the transformation of the public sphere through platformization, it was found that algorithms and design decisions operate as gatekeepers that structure visibility, reach, and the speed of communication. Analytically, the study adopts the category of dangerous speech, centered on effects and context (speaker, audience, message, medium, and environment), to assess risk and guide more nuanced responses to potentially violent online expressions. Based on the analysis of Instagram's measures, the study criticizes the low operational density of its guidelines against hate speech and the tension between maintaining platform integrity and maximizing engagement incentives. It argues that the most effective solution requires procedural guarantees of transparency, justification, appeal, and auditing, as well as the governance of information flows through graduated measures. In addition, it highlights the importance of long-term investments in digital education, particularly through media literacy and algorithmic literacy, as strategies to reduce susceptibility and the cumulative harms caused by dangerous speech on digital platforms. The article concludes that moderation is necessary but insufficient when applied in isolation. A polycentric arrangement that integrates law, platform engineering, and civic education is essential to adequately address hate speech in a context of ongoing transformation of collective communication.

Keywords/Palabras-claves/Mots-clés: Hate speech, Dangerous speech, Algorithmic governmentality, Content moderation, Digital education

1 INTRODUÇÃO

A expansão das redes sociais e, mais amplamente, dos ecossistemas digitais de comunicação e informação reconfigurou profundamente a esfera pública, transformando a maneira como indivíduos interagem e compartilham informações. O surgimento das plataformas sociais *on-line*, verdadeiras praças digitais, reduziu as barreiras de entrada para a fala pública de maneira tal que, o que antes dependia de *gatekeepers* tradicionais para ser veiculado – como editoras, emissoras e redações – passou a ser orquestrado por arquiteturas técnico-econômicas cuja lógica combina métricas de engajamento, dados comportamentais e modelos de recomendação, de modo tal que a conversação privada entre os indivíduos se dá, hoje, em infraestruturas privadas, com suas regras e modos de visibilidade próprios.

Nesses ambientes, a transformação da comunicação também se deu num nível temporal e escalar: a escala da comunicação digital é global, com um alcance jamais visto, e com uma temporalidade baseada em virais e picos rápidos, com postagens que atravessam comunidades em minutos, antes mesmo que agentes – humanos ou artificiais – consigam reagir para verificar, replicar ou moderar esses conteúdos. Isso posto, as mesmas propriedades que democratizam a fala também fomentam fricções, ao passo em que economias de atenção premiam o extraordinário, o confrontacional e o emocionante; e as métricas de sucesso dessas plataformas, nessa combinação de escala e velocidade, inadvertidamente acabam priorizando materiais polarizadores e limítrofes, favorecendo a proliferação da desinformação e de um discurso perigoso e potencialmente violento no ambiente digital.

Nas plataformas digitais, a forma como os usuários compartilham e interagem com os conteúdos *on-line* não é neutra, pois os algoritmos determinam as regras de visibilidade, as quais tornam certos tipos de mensagem mais propensos à circulação, numa curadoria automatizada ancorada em previsões de atenção. Aqui, e agudiza a tensão clássica entre liberdade de expressão e a manutenção de um ambiente *on-line* saudável: de um lado, a liberdade de expressão protege o dissenso, o teste de ideias impopulares e o desconforto inerente à vida democrática; e, de outro, um ambiente saudável demanda previsibilidade de regras, proteção contra assédio organizado e campanhas de dano e desumanização.

Chega-se, então, ao problema normativo-prático que orienta este estudo: dada a lógica de plataformas e a plasticidade contextual das comunicações *on-line*, a moderação de conteúdo realizada isoladamente pelo Instagram é suficiente para enfrentar a recorrência de discursos de ódio ou perigosos? Em outras palavras, é plausível esperar que um conjunto de regras e automatismos, operando sobretudo no nível do conteúdo, consiga mitigar riscos que emergem

do acoplamento entre incentivos de produto, dinâmicas de rede e comportamentos coletivos? A partir da análise das ferramentas adotadas pelo Instagram para operar contra esses conteúdos potencialmente violentos, a hipótese que se coloca neste trabalho é que, sem medidas que combinem transparência, responsabilização procedimental e formação crítica dos usuários, a moderação isolada, embora necessária, é insuficiente para preservar, no ambiente *on-line*, simultaneamente, a sua integridade e a vitalidade da liberdade de expressão.

A presente pesquisa adota uma metodologia qualitativa, de natureza teórico-normativa, estruturada em revisão bibliográfica e análise documental com abordagem dedutiva, num percurso metodológico que passa pelo mapeamento conceitual das categorias centrais de análise (como discurso de ódio e discurso perigoso) e pelo exame crítico de marcos normativos e documentos institucionais relevantes além das diretrizes e políticas do Instagram disponíveis publicamente, enquanto *corpus* documental. Assim, foi possível examinar o conceito epistemológico e jurídico do discurso perigoso e dos crimes de ódio no direito brasileiro, os parâmetros de moderação de conteúdo empregados *on-line* pelo Instagram, a fim de discutir o conceito de governamentalidade algorítmica como chave interpretativa da atuação das plataformas digitais e, daí, formular uma síntese argumentativa, visando identificar convergências, tensões e lacunas regulatórias, avaliando o potencial da educação digital como política de prevenção e mitigação de práticas intolerantes no ambiente *on-line*.

2 O ALGORITMO E A ECONOMIA POLÍTICA DAS PLATAFORMAS DIGITAIS: TENSÕES NA DIMENSÃO COLETIVA DA COMUNICAÇÃO *ON-LINE*

A emergência do digital não significou apenas a adoção de novas ferramentas, mas a reconfiguração das próprias condições de produção de conhecimento, coordenação social e estabilização normativa. Segundo Campos (2022, p. 257), "o digital transforma não apenas a geração de conhecimento social, mas também as interações e experiências mais íntimas de indivíduos uns com os outros e com as instituições, que influenciam de forma decisiva suas trajetórias sociais", de modo que não há que se falar, hoje, em um mundo digital separado da realidade analógica, mas em uma transposição do mundo real para o virtual. Nesse sentido, nas últimas décadas, com a popularização da internet e a expansão massiva de serviços orientados a dados, o direito – historicamente ancorado em institucionalidades estatais e na estabilização de expectativas – passou a operar sob novas pressões epistêmicas e temporais: a simultaneidade de eventos e a aceleração informacional desafiam mecanismos de garantia e de previsibilidade, obrigando o direito a se adaptar à autotransformação da sociedade digital.

Nesse novo regime jurídico, as plataformas digitais não são apenas uma infraestrutura técnica: são formas de organização da vida social que integram, num mesmo ambiente, captação de dados, mecanismos de recomendação e protocolos de visibilidade. Tal integração projeta efeitos não só sobre como circulam conteúdos, mas sobre o próprio modo como se produz valor e autoridade na esfera pública, de modo que a arquitetura de dados, os modelos algorítmicos e a aprendizagem de máquina se tornam elementos constitutivos do arranjo econômico e jurídico de plataformas, com implicações para a distribuição de poder comunicacional.

Se, no século XX, a relação entre meios de difusão (rádio, imprensa e TV) e grandes organizações estruturou a publicidade social, hoje a infraestrutura tecnológica global das plataformas desloca os eixos de coordenação e de legitimação. Não se trata de um vazio normativo, mas de uma nova combinação entre elementos horizontalizantes (redes, participação) e verticalizantes (governança algorítmica, termos privados), que tanto pode abrir oportunidades de liberdade quanto intensificar riscos de violação de direitos – e no cerne dessa transformação está a ocupação do espaço editorial por mecanismos de moderação automatizada.

Todavia, há uma tensão estrutural entre a promessa emancipatória das plataformas e seus mecanismos de ordenação invisível. As escolhas de *design* e o algoritmo operam como regulação de fato, instituindo regras de visibilidade e de alcance que moldam a conversação pública antes mesmo da intervenção jurídica formal; e, como resultado, se tem um duplo sistema de estruturação do espaço comunicativo, em que coexistem, de um lado, normas públicas que ainda partem de categorias de conteúdo, e, de outro, um complexo de decisões tecnológicas e contratuais que governam fluxo, descoberta e engajamento em escala transnacional. Para o direito, questão metodológica e normativa é como rearticular garantias de liberdade e de igualdade comunicativa num ambiente cuja temporalidade é a dos picos virais e das simultaneidades que escapam ao controle institucional clássico, pois essa nova economia política das plataformas digitais reconfigura a própria gramática de intervenção, deslocando o eixo do pós-fato – sanção – para o *ex ante* – a governança de fluxo. Em suma, a economia política das plataformas digitais exige que a dimensão coletiva da comunicação, como condição de possibilidade da própria liberdade de expressão, seja pensada para além das categorias tradicionais, cabendo o desafio de instituir, por desenho institucional e procedimental, contrapesos que preservem a vitalidade do debate público sem abdicar da proteção contra danos sistêmicos que decorrem da própria lógica de dados e de recomendação que tornou as plataformas socialmente indispensáveis (Campos, 2022).

3 CONCEPÇÃO EPISTEMOLÓGICA E JURÍDICO-NORMATIVA DO DISCURSO DE ÓDIO NO DIREITO BRASILEIRO

Partindo do pressuposto foucaultiano de que não há discurso dado de antemão, puro ou transparente (Foucault, 2011), o que se costuma denominar por discurso de ódio é um objeto construído no interior de regularidades discursivas, por meio de recortes, séries e critérios que selecionam o que conta como enunciável, verdadeiro, lícito ou ilícito. O discurso de ódio não é apenas uma categoria jurídica; é um ponto de convergência de diferentes formações discursivas, que estabilizam objetos, enunciados e regimes probatórios: trata-se de descrever a formação dos objetos e estratégias que dão inteligibilidade ao ódio como problema do direito, sem naturalizá-lo (Foucault, 2008).

À luz desse duplo enquadramento – o Estado como ordem jurídico-normativa e, simultaneamente, como forma de organização social –, se faz necessário deslocar o foco do questionamento acerca do que diz a lei: deve-se investigar, na verdade, como as relações sociais produzem e fazem operar categorias como discurso de ódio na esfera pública contemporânea, fortemente mediada por plataformas digitais. Isso significa acompanhar os procedimentos de seleção e rarefação pelos quais certos enunciados ganham estatuto de risco, passam a acionar expectativas de tutela, e, ao final, são institucionalizados em decisões, políticas e práticas administrativas. É precisamente nesse encontro entre formas jurídicas e dinâmicas sociológicas que se compreende por que o chamado discurso de ódio assume contornos fluidos e ambíguos - e porque, no Brasil, a proteção constitucional da liberdade de expressão convive com limites proporcionais quando a fala se converte em instrumento de violação de direitos.

No plano sociológico, o discurso de ódio circula em múltiplos campos e é utilizado por diferentes atores sociais, o que lhe confere contornos fluidos e, por vezes, ambíguos (Brown, 2017). Embora a liberdade de expressão ocupe posição nuclear enquanto direito fundamental consagrado na Constituição da República (Brasil, 1988), ela não é ilimitada, pois a tradição do constitucionalismo democrático no Brasil repudia a existência de direitos absolutos – com a excepcionalíssima ressalva da proibição da tortura (Bobbio, 2004). Como consequência, quando a expressão se converte em instrumento de violação de direitos, seja por incitar a violência, por promover discriminações ou por corroer as condições mínimas de participação pública de grupos vulneráveis, se abre espaço para restrições justificadas e para a responsabilização nas esferas penal, cível e administrativa.

No direito brasileiro não há uma categoria tipificada de modo unitário como crime de ódio, se tratando, na verdade, de um rótulo descritivo usado por pesquisas criminais e de direitos

humanos para designar condutas motivadas por hostilidade a atributos identitários de caráter odioso e discriminatório – de maneira tal que, por vezes, as manifestações captadas por essa ampla conceituação acabam carecendo de uma maior precisão conceitual e delimitação analítica e operacional. No entanto, cumpre salientar que a ausência normativa específica não significa inexistência de tutela jurídica, e o ordenamento brasileiro abriga mecanismos que alcançam condutas odiosas como racismo, xenofobia e homotransfobia (Tanure, 2021). No plano jurídico-positivo, a resposta estatal às manifestações usualmente rotuladas como crimes de ódio é construída a partir de um arcabouço normativo disperso, no qual se sobressai, notadamente, a Lei n.º 7.716/1989, cujo artigo 1º estabelece a punição de crimes resultantes de discriminação ou preconceito de raça, cor, etnia, religião ou procedência nacional (Brasil, 1989).

A esse núcleo de proteção jurídica se somam figuras penais correlatas do Código Penal, como incitação ao crime, ameaça e a injúria racial – esta última substancialmente reconfigurada por reformas legislativas recentes –, além de diplomas especiais de tutela antidiscriminatória (Brasil, 1940). O resultado é um mosaico que, embora não ofereça uma tipificação unitária de crime de ódio, como referido, permite enquadrar juridicamente condutas motivadas por hostilidade a atributos identitários. Tal desenho possui virtudes e custos: por um lado, ao invés de uma cláusula penal genérica, o ordenamento oferece tipos específicos e vias sancionatórias diversificadas, oferecendo uma tutela mais ampla; entretanto, esse justamente pode ser interpretado como um revés, pois dispositivos dispersos e, por vezes, redundantes podem dificultar a aplicação concreta do direito, sobretudo em cenários que misturam o *on-line* e o *off-line*, com múltiplos agentes e circulação transnacional de conteúdo.

Nesse contexto, ganha relevância, também, o Marco Civil da *Internet* (Brasil, 2014), que reordenou responsabilidades no ecossistema digital, fixando parâmetros de direitos dos usuários, neutralidade de rede, proteção de dados, e, sobretudo, uma arquitetura de responsabilização de intermediários. Em regra, a remoção de conteúdo depende de ordem judicial, com exceções específicas: o autor do ilícito responde diretamente nas esferas penal e cível; e os provedores têm deveres procedimentais de guarda de registros e de colaboração. Em termos práticos, isso significa deslocar o foco da punição imediata para processos de identificação, avaliação e decisão, um desenho compatível com a complexidade técnica e com garantias de devido processo informacional (Koche Júnior; Santos; Toledo, 2024).

Ainda no plano cível, o ordenamento oferece outros instrumentos de resposta ágil e graduada: além dos artigos 186 e 187 do Código Civil (Brasil, 2002), que punem, respectivamente, o ato ilícito e o abuso de direito – este último particularmente útil quando a forma expressiva é desvirtuada por violação de boa-fé, bons costumes e finalidade social –,

ganha relevo o dever de indenizar previsto no artigo 927, articulado ao critério de extensão do dano inscrito no artigo 944: esses dispositivos normativos permitem que a resposta seja proporcional ao gravame efetivamente produzido, inclusive quando o dano é extrapatrimonial e difuso, sem confundir crítica dura com desumanização.

No eixo tutelar, o Código de Processo Civil (Brasil, 2015) oferece instrumentos aptos a prevenir ou fazer cessar lesões: as tutelas inibitórias e as tutelas de urgências, previstas, respectivamente, nos artigos 497 e 300, permitem ao magistrado, inclusive em modalidade antecipatória, interromper a continuidade do dano, estancar a replicação de conteúdo e evitar reiterações enquanto se instrui o mérito; e em casos com prova pré-constituída, a tutela de evidência, prevista pelo artigo 311, também pode ser manejada para superar a inércia decisória típica do ambiente digital. Além disso, a eficácia desses instrumentos pode ser reforçada pelo instituto das astreintes, que pode funcionar como incentivo ao cumprimento célere de ordens de remoção despriorização ou desativação de perfis e conteúdos *on-line*.

Essas respostas jurídicas não silenciam o dissenso. Seu propósito é reconstituir condições mínimas de convivência discursiva, coibindo o rebaixamento de status e a expulsão simbólica de grupos do espaço público. Essa engenharia cível dialoga com a arquitetura técnico-organizacional das plataformas, ao passo em que essas medidas podem atuar nos fluxos digitais para ajustar recomendações, restringir alcance automatizado, rotular conteúdo para contextualização e reforçar o devido processo interno de regulação das plataformas digitais. Trata-se de procedimentalizar as liberdades, mantendo o direito de falar sem transformá-lo em direito à amplificação, quando a mensagem cruza limiares de risco identificáveis.

Desse panorama aqui exposto, emerge a dupla função estatal no espaço digital, enquanto ente incumbido da proteção da dignidade humana: de um lado, cabe ao Estado desfazer o senso comum de que a *internet* seria fora do alcance da ordem jurídica, afirmando a continuidade de garantias e deveres; e, de outro, possui ele o dever de tutelar de modo efetivo e proporcional os direitos de indivíduos e grupos atingidos, combinando educação midiática, mecanismos acessíveis de denúncia e recurso, transparência procedimental e sanções quando necessárias (Silva *et al.*, 2011) – pois a boa política jurídica para o enfrentamento do ódio não se esgota na punição, mas exige governança, *accountability* e alfabetização cívico-informacional, elementos indispensáveis para que a liberdade de expressão e a integridade do ambiente comunicativo se reforcem mutuamente, em vez de se anularem.

Nesse sentido, à luz de uma perspectiva que recusa essencialismos, importa tratar do conceito de discurso de ódio como uma construção analítica e normativa cuja inteligibilidade depende de contexto, assimetria de poder e efeitos projetados na esfera pública. Benesch (2023)

traz dois elementos cruciais para compreender tanto o discurso de ódio quanto categorias a ele correlatas, como o discurso tóxico, extremo e perigoso: em primeiro ponto, ressalta que se trata não de um fenômeno inato, mas ensinado e aprendido; e, em segundo lugar, lembra que seus impactos tendem a ser cumulativos, se consolidando por repetição e reforço social. Nessa chave, o foco se desloca de ofensas pontuais para padrões e dinâmicas que corroem, no tempo, as condições de participação de determinados grupos, de maneira consoante ao entendimento de Waldron (2010), que sublinha que a relevância jurídica do discurso de ódio exige exteriorização do pensamento, ou seja, a passagem de ideias a uma linguagem publicada – seja verbal, visual ou performativa. O ponto, aqui, não é policiar pensamentos, mas avaliar enunciados e práticas comunicativas quando projetam danos sociais, como a intimidação difusa, desumanização de minorias, erosão de *status* e expulsão simbólica do espaço público.

Para distinguir o que é mera grosseria do que ganha relevo jurídico, convém delimitar alvo, conteúdo e marcadores de risco. Nesse sentido, Meyer-Pflug (2009) define o discurso de ódio como a externalização de concepções preconceituosas dirigidas a indivíduos ou grupos, e, numa linha semelhante, Brugger (2007, p. 118) o caracteriza como manifestações que "[...] tendem a insultar, intimidar ou assediar pessoas em virtude de sua raça, cor, etnicidade, nacionalidade, sexo ou religião ou que tem a capacidade de instigar violência, ódio ou discriminação contra tais pessoas". De modo convergente, diretrizes internacionais, como as da União Europeia (2018) ressaltam que uma simples expressão de desprezo ou violência não basta para configurar discurso de ódio: exige-se, em geral, a verificação de um ataque concreto a uma pessoa ou a um grupo por uma característica identitária partilhada. Nesse sentido, a Organização das Nações Unidas (2019, p. 2), para evitar exclusões arbitrárias, propõe uma lista não exaustiva de critérios – como religião, etnia, nacionalidade, raça, cor, ascendência, gênero, entre outros –, reconhecendo a variabilidade contextual dos alvos.

Aqui, no entanto, surge a primeira controvérsia: quem é reconhecido como potencial alvo? A literatura registra ausência de consenso e alerta contra os riscos de um alargamento indevido do conceito de discurso de ódio e de seu estreitamento artificial. No plano analítico, portanto, a motivação discriminatória é peça importante, pois confere gravidade ao projetar o dano para além da vítima individual, atingindo o grupo e o ambiente. O desafio prático-jurídico, no entanto, está em como reconhecer e provar essa motivação: deve-se se a reduzi-la a um elemento subjetivo de difícil constatação ou deve ser admitido inferi-la a partir do contexto social em que ocorre a violência? Benesch (2023). Reduzir a motivação extremista e discriminatória a um elemento subjetivo explícito torna sua demonstração inviável em muitos casos, e ignorá-la, por outro lado, dilui o sentido antidiscriminatório da tutela.

O caminho razoável parece ser admitir a inferência contextual, a partir de padrões de linguagem e histórico do emissor, além do alvo em concreto e do momento sociopolítico em que foi proferido o discurso, compondo, assim, um quadro probatório que respeite garantias sem blindar retóricas nocivas sob o manto da ambiguidade. É nesse ponto que a proposta de discurso perigoso de Benesch (2023) oferece utilidade preventiva e operativa: a autora indaga se o discurso de ódio deve, necessariamente, promover ódio ou se bastaria que alguém se sentisse atingido por ele; e questiona se deve ser levada em consideração somente a intenção do autor ou os efeitos do discurso produzidos sobre os destinatários, questionando a intensidade, duração e violência necessárias para que uma emoção seja juridicamente qualificada como ódio discursivo. Ao invés de se fixar no conteúdo intrínseco ou na intenção subjetiva, a categoria centraliza o efeito do discurso, considerando perigoso aquela expressão que aumenta o risco de violência ou discriminação contra um grupo, a partir de certos contextos e variáveis.

> *Dangerous speech is a narrower and more precisely bounded category than hate speech, the most prevalent term in academic literature and common discourse* [...]. *The term hate speech itself presents important questions that have not yet been consistently answered. First, must hate speech express hatred, promote hatred, or make someone feel hated?* [...] *If it is the intention of the author that is definitive, the state of another person's mind is not always easy to discover, especially when its expression is found online* (Benesch, 2023, p. 187).

A autora, então, sistematiza cinco dimensões para uma aferição graduada do que seria um discurso perigoso: (i) o orador, sua autoridade, credibilidade, e acesso a coerção; (i) a audiência, suas vulnerabilidades, estado emocional e exposição prévia a narrativas hostis; (iii) a mensagem, a partir de seus marcadores retóricos, como desumanização e retórica de contaminação; (iv) o contexto, levando em consideração tensões sociais, gatilhos históricos e deslegitimação de salvaguardas; e (v) os meios ou a mídia, com base no alcance, velocidade e exclusividade informacional do canal (Benesch, 2023, p. 188-191).

Nesse sentido, dois corolários daí decorrem: em primeiro lugar, não basta a possibilidade abstrata de incitar violência, sendo necessário avaliar o risco concreto à luz de todas as variáveis conjuntamente; além disso, a repetição importa, de modo que exposições reiteradas naturalizam ideias hostis e dessensibilizam audiências. Em ecossistemas digitais de alta velocidade e recomendação algorítmica, essa dinâmica cumulativa se exacerba, pois a arquitetura de distribuição pode transformar sinais fracos em normas de sentido, corroendo os limiares sociais que inibem danos. Essa perspectiva, aliás, dialoga com o pensamento foucaultiano, segundo o qual os objetos construídos pelo discurso estão em permanente formação e transformação (Silva; Machado Júnior, 2015).

A ausência de consenso acadêmico sobre o que constitui discurso de ódio reforça a pertinência dessa crítica, e a análise dessas variáveis oferece maior objetividade para a compreensão da violência *on-line*, fornecendo subsídios relevantes tanto para a responsabilização civil e penal quanto para a formulação de parâmetros de moderação de conteúdo em plataformas digitais. O conceito teórico de discurso perigoso possui, como se vê, uma possibilidade de aplicação preventiva maior e mais específica que a ampla noção de discurso de ódio, comum na literatura: considera-se perigoso o discurso que contribui para a criação de um ambiente favorável à violência contra determinado grupo, de modo que a centralidade recai, portanto, sobre o resultado do discurso (Benesch, 2023).

Assim, a categoria de discurso perigoso cumpre exatamente a função preventiva que este trabalho demanda, nascendo da detecção de padrões retóricos recorrentes em falas de líderes políticos, religiosos e culturais antes de erupções de violência em distintos tempos e lugares (Benesch, 2023, p. 188), tendo o propósito de democratizar esses conhecimentos para que sociedades consigam reconhecer os sinais e interromper a escalada do dano. Ao deslocar o foco para efeitos e contexto, e operacionalizar a análise em cinco dimensões (orador, audiência, mensagem, contexto e meios), o modelo permite aferir risco em gradações, capturar a cumulatividade da exposição e orientar respostas proporcionais no direito e na governança de plataformas, sem sacrificar o núcleo da liberdade de expressão.

4 DISCURSO DE ÓDIO E DISCURSO PERIGOSO: TENSÕES REGULATÓRIAS NO AMBIENTE DIGITAL VERIFICÁVEIS NO CASO DO INSTAGRAM

A passagem da esfera pública das organizações para a esfera pública das plataformas, como conceitua Campos (2022), reconfigura quem define o que se torna visível, a que ritmo circula e com quais efeitos sociais. As plataformas digitais funcionam como *gatekeepers* algorítmicos: mediadores que organizam interações por meio de dados, modelos de recomendação e protocolos de visibilidade – e, nesse cenário, o caso do Instagram é exemplar por condicionar a circulação de conteúdo por meio de decisões de *design* e de mediação algorítmica, ou seja, por decisões que não apenas espelham o social, mas coproduzem novas estruturas de sociabilidade e de atenção. No entanto, essa plataformização da comunicação não opera em vazio normativo: ela se ancora em uma infraestrutura jurídica transnacional que impulsiona a autorregulação e desloca parte da governança para procedimentos internos, o que exige procedimentalizar direitos comunicativos para recompor garantias na era digital:

> Essa adaptação regulamentar revela-se especialmente importante no atual cenário, onde o novo meio, que é indiscutivelmente um amplificador de possibilidades, coincide em grande parte com o modelo de negócios de algumas poucas empresas. [...] Neste contexto, os novos intermediários têm amplo controle de fato do acesso à nova dimensão coletiva da comunicação e podem, assim, decidir em larga medida sobre o exercício efetivo das liberdades de terceiros e, em última instância, sobre a formação da nova dimensão coletiva da comunicação, com impactos claros na constituição da esfera pública democrática. Uma metodologia jurídica adequada à plataforma deve, portanto, tomar necessariamente como ponto de partia esta nova factualidade da sociedade da plataforma (Campos, 2022, p. 314).

A distinção entre discurso de ódio e discurso perigoso ganha saliência no Instagram porque a maior parte das decisões acontece *ex ante* por mecanismos automatizados e fluxos de trabalho terceirizados, antes mesmo de qualquer juízo judicial. Essa mediação técnico-organizacional tem uma racionalidade própria, descrita por Rouvroy e Berns (2015) como governamentalidade algorítmica, revisitando o conceito foucaultiano:

> A governamentalidade algorítmica não produz qualquer subjetivação, ela contorna e evita os sujeitos humanos reflexivos, ela se alimenta de dados 'infraindividuais' insignificantes neles mesmos, para criar modelos de comportamento ou perfis supraindividuais sem jamais interpelar o sujeito, sem jamais convocá-lo a dar-se conta por si mesmo daquilo que ele é, nem daquilo que ele poderia se tornar (Rouvroy; Berns, 2015, p. 42).

No entanto, é certo que a moderação de conteúdos abusivos em um ambiente descentralizado e dinâmico como o da *internet* não é tarefa simples (Tenorio; Moreira, 2023). A moderação em plataformas envolve cadeias globais de terceirização e padrões de opacidade contratual, o que dificulta o escrutínio público e a validação independente de critérios. Além disso, Benesch (2023) explica que a remoção de conteúdos, isoladamente, é uma estratégia pouco efetiva para combater processos de radicalização social, sendo preferível a adoção de estratégias como campanhas de alfabetização digital do público – incentivando-o a refletir antes de externalizar o seu pensamento – e mais refinados processos de redução da amplificação algorítmica quando o conteúdo versar sobre discursos perigosos:

> *In response to public pressure and legal requirements, especially those with social media platforms, tech companies are increasingly trying out new techniques to try to more effectively identify and deal with hate speech, dangerous speech, and other harmful content on their online turf. While removing such content is the most visible remedy, it is a heavy-handed approach, and there are many alternatives that better protect freedom of expression but need to be better understood, including downranking content (reducing its algorithmic amplification), 'nudging' users to reconsider their words before they are posted* […] *(Benesch, 2023, p. 193).*

A noção de governamentalidade algorítmica ajuda a entender por que o problema não se esgota em listas de conteúdos proibidos. Como descrevem Rouvroy e Berns (2015), a

modelagem por correlações termina por colonizar o espaço público com uma esfera privada hipertrofiada, filtrando informação, radicalizando opiniões e corroendo a experiência comum, tudo isso sob a pressão da economia da atenção. Não obstante, a autorregulação é instrumento indispensável para assegurar um ambiente de debate público equilibrado e para preservar os direitos de todos os usuários das plataformas digitais. Trata-se, contudo, de uma moderação que não deve ser arbitrária ou subjetiva, mas pautada em critérios objetivos e diretrizes transparentes, previamente disponibilizados aos usuários e acompanhados da possibilidade de exercício do contraditório e da ampla defesa (Tenorio; Moreira, 2023). No caso do Instagram, moderar não se limita a retirar determinado conteúdo compartilhado *on-line*; é, em fato, governar fluxos: a resposta regulatória mais promissora desloca o foco da criação *ad hoc* de novos ilícitos para exigências procedimentais verificáveis, com relatórios que explicitem parâmetros de recomendação e despriorização, justificativas compreensíveis para medidas de visibilidade, mecanismos de recurso com taxas de reversão auditáveis, e testes de impacto sobre grupos vulneráveis.

Uma regulação eficaz não pode se limitar ao binarismo de permitir ou proibir conteúdos, nem se reduzir a listas estáticas de termos. Em ambientes como o Instagram, cujo desenho técnico multiplica formatos, velocidades e circuitos de descoberta, o eixo decisório se desloca para como o fluxo é organizado: quem vê o quê, em que ordem, com que ênfase e com que fricções. A opacidade estrutural dos processos internos das plataformas digitais exige contrapesos institucionais, e, nesse sentido, além de cláusulas proibitivas, é indispensável procedimentalizar direitos comunicativos, com previsibilidade de regras, notificação ao usuário afetado, motivação compreensível das decisões, canais de recurso efetivos e mecanismos de auditoria periódica. A autorregulação, nessa chave, não substitui o Estado; ela pressupõe um arranjo policêntrico, em que autoridades públicas e a sociedade civil definem padrões verificáveis a serem implementados pelas plataformas digitais. No Brasil, esse desenho dialoga com o Marco Civil da *Internet* (Brasil, 2014) e pode ser informado pela garantia constitucional da proteção de dados, sem colidir com a liberdade de expressão (Brasil, 1988).

O Instagram, hoje, oferece essencialmente dois instrumentos para regular o conteúdo publicado em sua plataforma: os termos de uso e as diretrizes da comunidade. Ambos contêm cláusulas proibitivas sobre violência, terrorismo, crime organizado e grupos de ódio, bem como enunciam a remoção de ameaças concretas de violência, discurso de ódio e perseguição com base em características protegidas, como raça, etnia, nacionalidade, sexo, gênero, identidade de gênero, orientação sexual, religião, deficiência ou doença (Perguntas, 2018). Em materiais mais recentes, a empresa controladora do Instagram refina a definição de conduta de ódio adotada

pela plataforma digital como ataques diretos a pessoas por características protegidas, além de proteção diferenciada para grupos como refugiados e migrantes (Meta, 2025).

> Definimos conduta de ódio como ataques diretos a pessoas, e não a conceitos e instituições, baseado no que chamamos de características protegidas: raça, etnia, nacionalidade, deficiência, religião, casta, orientação sexual, sexo, identidade de gênero e doença grave. Além disso, consideramos a idade uma característica protegida quando referida juntamente com outra característica também protegida. [...] Às vezes, com base em particularidades locais, consideramos palavras ou frases específicas como proxies usados com frequência para grupos com características protegidas. [...] Removemos discursos desumanizantes, alegações de imoralidade ou criminalidade grave e insultos. Também removemos estereótipos prejudiciais, que definimos como comparações desumanizantes historicamente usadas para atacar, intimidar ou excluir grupos específicos e que muitas vezes estão ligadas à violência no meio físico. Por fim, removemos insultos graves, expressões de desprezo ou repulsa, xingamentos e incitação à exclusão ou segregação quando direcionados a pessoas com base em características protegidas (Meta, 2025, n. p.)

O problema central, no entanto, não reside na declaração de princípios, mas na sua operacionalização, pois essas diretrizes permanecem genéricas e pouco sensíveis ao contexto sociotécnico específico da plataforma. Dentre as principais plataformas *on-line*, o Instagram é a que menos se aprofunda, em suas normas internas, a respeito da temática do discurso de ódio, de modo que suas diretrizes apenas tratam, de forma superficial, as dinâmicas relacionadas à circulação de discursos de ódio, vinculando os usuários às políticas de sua empresa controladora: ao invés de um *playbook* específico para seu ecossistema, o serviço herda formulações amplas da Meta e não traduz adequadamente as especificidades de formato, comunidade e práticas próprias do Instagram (Santos *et al.*, 2023). Falta densidade normativa para responder às variações por formato, por circuito de circulação e por dinâmica de rede; e, mesmo quando a plataforma reconhece *proxies* locais, não explicita critérios verificáveis de aplicação. "A falta de especificidade da plataforma fortalece a dificuldade na moderação de conteúdo de ódio, uma vez que suas diretrizes não levam em conta as especificidades de formato, comunidade, gramáticas e práticas do Instagram" (Santos *et al.*, 2023, p. 10).

Essa lacuna de operacionalização aparece, sobretudo, onde mais importa: nas políticas de fluxo. As regras dizem o que é proibido, mas pouco esclarecem como o Instagram atenua a circulação de conteúdos limítrofes, e, sem indicadores auditáveis a moderação permanece opaca e reativa, quando deveria ser preventiva e proporcional. Por fim, permanece a tensão estrutural entre integridade do ambiente e incentivos de engajamento, pois tal cenário também revela uma conexão entre o crescimento dos discursos de ódio *on-line*, a intensificação da polarização social e a utilização desse tipo de retórica por políticos, influenciadores e canais de mídia diversos como estratégia para gerar engajamento (Koche Júnior; Santos; Toledo, 2024).

Retóricas polarizadoras e conteúdos limítrofes tendem a performar melhor em métricas de atenção, de modo que a própria plataforma, de maneira indireta, se beneficia economicamente da propagação e da viralização de conteúdos abusivos e violentos.

Para que a moderação seja intelectualmente sólida e operacionalmente eficaz, as plataformas precisam dialogar com a melhor evidência científica disponível sobre o discurso perigoso e o discurso de ódio, e sobre as dinâmicas de risco *on-line* – tanto para a qualidade do ambiente informacional quanto para a gestão de responsabilidades inerentes ao negócio. O ponto de partida, portanto, não é apenas listar proibições, mas incorporar um referencial de risco sensível a contexto, cumulatividade e assimetrias de poder (Benesch, 2023). Nesse arranjo, o controle de condutas no digital não se esgota em normas abstratas: ele passa, decisivamente, pelo próprio código e pela arquitetura do serviço – aquilo que os usuários podem fazer, com que fricções, a que velocidade e com qual alcance (Tenorio; Moreira, 2023).

Em termos clássicos, Lessig (1999) já demonstrava que o desenho técnico regula tanto quanto a lei escrita; e que a *internet* pode ser regulada tanto pela alteração de sua arquitetura e de seu código – que moldam a forma como os indivíduos interagem – quanto pela criação de normas jurídicas que considerem essas mudanças para alcançar efetividade. Logo, uma política responsável exige articulação entre mecanismos técnicos e normativos: ajustes de ranking, rótulos e limites de propagação combinados a regras claras, previsíveis e controláveis, aptas a enfrentar os riscos sociais e comunicacionais que emergem *on-line*. A moderação do Instagram deve ser entendida, assim, como parte de um esforço regulatório mais amplo, orientado por parâmetros claros, proporcionais e justificáveis que conciliem o núcleo da liberdade de expressão com a proteção da esfera pública contra manifestações perigosas – o que implica um devido processo informacional e o controle jurisdicional, sempre que ele se mostrar necessário – mas sem abdicar do aprendizado organizacional contínuo (Tenorio; Moreira, 2023).

No plano operacional, a literatura de discurso perigoso oferece uma base de classificação graduada para tanto, ao passo em que Benesch (2023) observa que, embora o contexto não possa ser plenamente detectado por sistemas automáticos, é possível usar modelos para triagem e detecção de padrões recorrentes, desde que mediante supervisão humana e com validação independente. No sentido do aperfeiçoamento da moderação, ela propõe critérios de classificação do grau de risco do discurso perigoso, e afirma que, "[...] *though the dangerousness of speech depends greatly on context, which cannot be detected and evaluated automatically, it may be possible to build classifiers for dangerous speech that operate by detecting similarities and patterns in it* (Benesch, 2023, p. 194)".

Na governamentalidade, o cidadão não se configura como sujeito passivo: ele também resiste, uma vez que sua prática e seu discurso produzem poder e saber. Contudo, a resistência à governamentalidade algorítmica revela-se mais complexa, pois o indivíduo, em grande medida, não se dá conta de que está sendo governado. Assim, as *big techs* passam a orientar condutas digitais e a definir, de forma quase imperceptível, o que deve ou não ser enquadrado como discurso de ódio.

Conforme Foucault (2008), não existem saberes dissociados de relações de poder. Desse modo, ao criarem suas próprias diretrizes acerca do que constitui discurso de ódio, plataformas como o Instagram revestem-se de um poder-saber praticamente ilimitado, caso não haja limitação imposta pelo Estado Democrático de Direito. Nesse cenário, evidencia-se a importância da produção acadêmica, da legislação e das decisões judiciais, que buscam tensionar e delimitar o campo do que pode ser considerado — ou rejeitado — como discurso odioso.

5 DA GOVERNAMENTALIDADE À PLATAFORMIZAÇÃO: EDUCAÇÃO DIGITAL COMO CONTRAPONTO AO DISCURSO PERIGOSO

Conforme Foucault (2008), não existem saberes dissociados de relações de poder; e não há exercício de poder sem produção de saber. É nesse ponto que a noção de governamentalidade algorítmica (Rouvroy; Berns, 2015) se torna elucidativa: diferentemente de uma regulação centrada em conteúdos tipificados, o governo pelos algoritmos opera com correlações infraindividuais, perfis probabilísticos e retroalimentação de atenção. Seu efeito não é interpelar sujeitos para que se reconheça numa norma, mas preordenar os contextos de visibilidade e descoberta, numa racionalidade que se apresenta como técnica, mas produz normatividade ao modelar o campo de possibilidades. Em ambientes como o Instagram, isso significa que a fronteira entre ódio e perigo é, em larga medida, decidida na governança do fluxo, antes de qualquer remoção (Benesch, 2023).

Se a governamentalidade é, em Foucault, ao mesmo tempo poder e conhecimento, a resposta democrática não é apenas multiplicar proibições: é reinscrever esses poderes em procedimentos públicos. Nesse sentido, o que hoje se denomina por educação digital deve ser pensado como política pública de governo das condutas, uma camada preventiva que incide sobre a formação de repertórios, hábitos de atenção e leitura de contexto. Afinal, enquanto as normas jurídicas atuam sobre as consequências dos atos, a educação possui o potencial de promover mudanças comportamentais efetivas, por ter suas esfera de atuação na raiz das

questões sociais, se apresentando como instrumento seguro para o aprimoramento das condutas humanas, tanto no campo jurídico quanto nas interações sociais (Abrusio, 2020).

O Relatório de Recomendações para o Enfrentamento ao Discurso de Ódio e ao Extremismo no Brasil organiza a resposta pública em seis eixos: (i) educação e cultura em direitos humanos; (ii) escola e universidade como promotoras de paz e convivência democrática; (iii) internet segura, educação midiática e comunicação popular/comunitária; (iv) proteção e reparação às vítimas; (v) dados e pesquisa para subsidiar ações e políticas; e (vi) boas práticas para jornalistas e comunicadores (Brasil, 2023). Lidos à luz da governamentalidade e da plataformização, esses eixos compõem pilares para uma aposta que não é punitiva, mas estrutural: fortalecer as capacidades individuais e institucionais para reduzir a periculosidade a montante. Ao invés de moralizar as plataformas digitais e os conteúdos *on-line*, o que se deve buscar é alfabetizar para o algoritmo e para a retórica: explicar como fluxos são ordenados, treinar o reconhecimento de marcadores de discurso perigoso e fortalecer competências de contexto. Nesse quadro, alfabetização midiática e literacia algorítmica ocupam lugar central. Não se trata apenas de informar o usuário, mas de formar competências para identificar manipulações, reconhecer padrões de incitação e compreender efeitos sociais das próprias publicações (Curvêlo; Angelim; Camargo, 2025).

A articulação com as propostas de Benesch (2023) permite dar operacionalidade preventiva a essa educação: ao invés de focar exclusivamente em intenção subjetiva ou listas de palavras, os programas didáticos devem treinar a leitura integrada das cinco variáveis do discurso perigoso (orador, audiência, mensagem, contexto e meio) e o reconhecimento de marcadores retóricos recorrentes. Nesse processo, não se trata apenas de fornecer informação, mas de desenvolver habilidades de leitura crítica que capacitem o cidadão a questionar e resistir a narrativas nocivas e a reagir com contramedidas proporcionais.

> Nesse cenário, a educação digital é uma ferramenta poderosa para empoderar os indivíduos, tornando-os mais resilientes às ameaças do ambiente online, como o discurso de ódio. Ao desenvolver habilidades críticas e promover o uso responsável da internet, ela contribui para a construção de uma sociedade mais informada, engajada e capaz de defender os valores democráticos e os direitos humanos no espaço digital (Curvêlo; Angelim; Camargo, 2025, p. 3249).

6 CONSIDERAÇÕES FINAIS

O presente trabalho partiu de um diagnóstico simples e exigente: a esfera pública contemporânea é coproduzida por arquiteturas sociotécnicas privadas que ordenam fluxos de

atenção e, com isso, redistribuem normatividades; e buscou analisar a suficiência da moderação de conteúdo do Instagram como medida isolada para prevenir a disseminação de discursos odiosos e perigosos, investigando as tensões entre as políticas das plataformas digitais, o conceito de governamentalidade algorítmica e o cenário jurídico brasileiro.

Do ponto de vista jurídico-positivo, o ordenamento brasileiro oferece um mosaico de tutelas (constitucionais, penais, civis e processuais) suficiente para enfrentar manifestações odiosas, desde que coordenado e aplicado com proporcionalidade. O estudo indicou que a ampliação casuística do direito penal é pouco desejável; e que a via mais promissora para o combate ao discurso de ódio *on-line* combina responsabilização civil, tutelas inibitórias ou urgentes e obrigações estruturais, articuladas a um devido processo informacional na camada privada que, de fato, organiza a circulação. Em outras palavras: mais do que criar novos tipos, é preciso procedimentalizar garantias onde o debate efetivamente acontece.

O exame do Instagram reforçou esse ponto: observou-se que as plataformas digitais, ao criarem suas próprias diretrizes sobre o que constitui discurso de ódio ou violento, exercem um poder-saber de difícil contestação, orientando condutas digitais de modo quase imperceptível. A incorporação do conceito de discurso perigoso oferece uma trilha operativa e preventiva, permitindo avaliar o risco através de cinco variáveis – o orador, a audiência, a mensagem, o contexto e o meio. Esse deslocamento do debate acerca da remoção do conteúdo para a governança de fluxos permite tratar a cumulatividade do dano sem sacrificar o núcleo da liberdade de expressão nem transformar o direito de falar em direito à amplificação.

Nesse sentido, a educação digital surge como política de longo prazo e baixo custo democrático: alfabetizar para o algoritmo e para a retórica, treinar leitura de contexto e reconhecimento de padrões, reduzir suscetibilidade e repetição que naturalizam hostilidade. Integrada à moderação e à governança, a educação melhora a qualidade das denúncias, diminui a demanda por remoções controversas e reconstitui condições mínimas de convivência discursiva – não moralizando a rede, mas fortalecendo a cidadania informacional.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRUSIO, Juliana. Os limites da liberdade de expressão na *internet*. **Revista Brasileira de Educação e Cultura**. São Gotardo, v. 11, n. 1, p. 76-97, jan./dez. 2020. Disponível em: https://www.periodicos.cesg.edu.br/index.php/educacaoecultura/article/view/507/0. Acesso em: 26 set. 2025.

BENESCH, Susan. *Dangerous speech*. *In*: ***Challenges and perspectives of hate speech research***. STRIPPEL, Christian. *et al* (Orgs.). 1. ed. Berlim: [*s. n.*], 2023. p. 185-197.

BOBBIO, Norberto. **A era dos direitos**. 1. ed. Tradução: Carlos Nelson Coutinho. Rio de Janeiro: Elsevier, 2004.

BRASIL. Congresso Nacional. **Lei nº 7.716, de 5 de janeiro de 1989**. Define os crimes resultantes de preconceito de raça ou de cor. Diário Oficial da União. Brasília, 6 de janeiro de 1989. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/L7716compilado.htm. Acesso em: 26 set. 2025.

BRASIL. Congresso Nacional. **Lei nº 10.406, de 10 de janeiro de 2002**. Institui o Código Civil. Diário Oficial da União. Brasília, 11 de janeiro de 2002. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Acesso em 26 set. 2025.

BRASIL. Congresso Nacional. **Lei nº 12.965, de 23 de abril de 2014**. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Diário Oficial da União. Brasília, 24 de abril de 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 26 set 2025.

BRASIL. Congresso Nacional. **Lei nº 13.105, de 16 de março de 2015**. Código de Processo Civil. Diário Oficial da União. Brasília, 17 de março de 2015. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm. Acesso em: 26 set 2025.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Diário Oficial da União. Brasília, 5 de outubro de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 26 set 2025.

BRASIL. Presidência da República. **Decreto-Lei nº 2.848, de 7 de dezembro de 1940**. Código Penal. Diário Oficial da União. Rio de Janeiro, 31 de dezembro de 1940. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: 26 set. 2025.

BRASIL. Ministério dos Direitos Humanos e da Cidadania. **Relatório de recomendações para o discurso de ódio e ao extremismo no Brasil**. 1. ed. Brasília: [*s. n.*], 2023.

BROWN, Alexander. *What is hate speech? Part 2: family resemblances*. **Law and Philosophy**. [*s. l.*], v. 36, n. 5, p. 561-613, out. 2017. Disponível em: https://link.springer.com/article/10.1007/s10982-017-9300-x. Acesso em: 26 set. 2025.

BRUGGER, Winfried. Proibição ou proteção do discurso do ódio? Algumas observações sobre o direito alemão e o americano. **Revista de Direito Público**. Brasília, v. 4, n. 15, p. 117-136, jan./mar. 2007. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/1418. Acesso em: 26 set. 2025.

CAMPOS, Ricardo. **Metamorfoses do direito global**: sobre a interação entre direito, tempo e tecnologia. 1. ed. São Paulo: Editora Contracorrente, 2022.

CURVÊLO, João Paulo de Sousa; ANGELIM, Gláucio de Aquino Cabral; CAMARGO, Maria Emilia. Educação digital e alfabetização midiática como ferramentas de combate ao discurso de ódio. **Revista Ibero-Americana de Humanidades, Ciências e Educação – REASE**. São Paulo, v. 11, n. 8, p. 3244-3254, ago. 2025. Disponível em: https://periodicorease.pro.br/rease/article/view/20889. Acesso em: 26 set. 2025.

FOUCAULT, Michel. **A arqueologia do saber**. 7. ed. Tradução: Luiz Felipe Baeta Neves. Rio de Janeiro: Forense Universitária, 2008.

FOUCAULT, Michel. **A ordem do discurso**: aula inaugural no Collège de France, pronunciada em 2 de dezembro de 1970. 21. ed. Tradução: Laura Fraga de Almeida Sampaio. São Paulo: Edições Loyola, 2011.

KOCHE JÚNIOR, Marcelo Ioris; SATOS, Mateus Arino dos; TOLEDO, Claudia Mansani Queda de. O direito à liberdade de expressão e o combate aos discursos de ódio no âmbito das redes sociais através da regulação pelo estado. **Revista Jurídica Unicuritiba**. Curitiba, v. 2, n. 78, p. 303-328, abr./jun. 2024. Disponível em: https://revista.unicuritiba.edu.br/index.php/RevJur/article/view/6662/pdf. Acesso em: 26 set. 2025.

LESSIG, Lawrence. The law of the horse: what cyberlaw might teach. **Harvard Law Review**. Cambridge, v. 113, n. 2, p. 501-549, dez. 1999. Disponível em: https://cyber.harvard.edu/works/lessig/LNC_Q_D2.PDF. Acesso em: 26 set. 2025.

MEYER-PFLUG, Samantha Ribeiro. **Liberdade de expressão e discurso de ódio**. 1. ed. São Paulo: Revista dos Tribunais, 2009.

MIRANDA JUNIOR, Gilberto. WILKE, Valéria Cristina Lopes. Institucionalidade, governamentalidade e inteligência artificial: democracia inteligente ou democracia artificial?. **LOGEION: Filosofia da informação**. Rio de Janeiro, v. 11, p. 1-20, nov. 2024. Disponível em: https://revista.ibict.br/fiinf/article/view/7373. Acesso em: 26 set. 2025.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. ***United Nations strategy and plan of action on hate speech***. [*s. l.*]: [*s. n.*], 2019. Disponível em: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf. Acesso em: 26 set. 2025.

PERGUNTAS frequentes sobre as diretrizes da comunidade do Instagram. [*s. l.*]: 2018. Disponível em: https://about.instagram.com/pt-br/blog/announcements/instagram-community-guidelines-faqs. Acesso em: 26 set. 2025.

ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação?. **Revista Eco Pós**. Rio de Janeiro, v. 18, n. 2, p. 36-56, 2015. Disponível em: https://revistaecopos.eco.ufrj.br/eco_pos/article/view/2662. Acesso em: 26 set. 2025.

SANTOS, Luiza Carolina dos. *et al*. Discurso de ódio *on-line*: uma análise das políticas das plataformas digitais para moderação de conteúdo. **E-Compós**. Brasília, v. 26, p. 1-22,

jan./dez. 2023. Disponível em: https://www.e-compos.org.br/e-compos/article/view/2709. Acesso em: 26 set. 2025.

SILVA, Giuslane Francisca da; MACHADO JÚNIOR, Sérgio da Silva. O discurso em Michel Foucault. **Revista Eletrônica História em Reflexão**. Dourados, v. 8, n. 16, [*n. p.*], jul./dez. 2014. Disponível em: https://ojs.ufgd.edu.br/historiaemreflexao/article/view/3821. Acesso em: 26 set. 2025.

SILVA, Rosane Leal da. *et al*. Discursos de ódio em redes sociais: jurisprudência brasileira. **Revista Direito GV**. São Paulo, v. 7, p. 445-468, jul. 2021. Disponível em: https://www.scielo.br/j/rdgv/a/QTnjBBhqY3r9m3Q4SqRnRwM/?lang=pt. Acesso em: 26 set. 2025.

TANURE, Augusto Lacerda. **A tolerância e a liberdade de expressão como possibilidade de cura do discurso de ódio**. 2021. Dissertação (Mestrado em Direito) – Programa de Pós-Graduação em Direito, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, 2021.

TENORIO, Caio Miachon. MOREIRA, Diogo Rais Rodrigues. Moderação de conteúdo pelas mídias sociais. **Revista Internacional Consinter De Direito**. Porto, ano 9, n. 17, [*n. p.*], 2023. Disponível em: https://revistaconsinter.com/index.php/ojs/article/view/438. Acesso em: 26 set. 2025.

META. Central de Transparência. **Conduta de ódio**. [*s. l.*], 2025. Disponível em: https://transparency.meta.com/pt-br/policies/community-standards/hate-speech/. Acesso em: 26 set. 2025.

UNIÃO EUROPEIA. Comissão Europeia. ***Countering illegal hate speech online***. Bruxelas, 2018. Disponível em: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/memo_18_262/MEMO_18_262_EN.pdf. Acesso em: 26 set. 2025.

WALDRON, Jeremy. *Dignity and defamation: the visibility of hate*. **Harvard Law Review**. Cambridge, v. 123, n. 7, p. 1597-1657, mai. 2010. Disponível em: https://harvardlawreview.org/archives/vol-123-no-7/. Acesso em: 26 set. 2025.