
**“DECIFRA-ME OU DEVORO-TE”: a moderação de conteúdo como forma de combate
ao discurso de ódio no ecossistema digital do Youtube**

**“DECIPHER ME OR I SHALL DEVOUR YOU”: content moderation as a tool to fight
hate speech on Youtube’s digital ecosystem**

Victoria Vital de Castro Macedo Costa¹

RESUMO

Este trabalho se propõe a investigar a questão do combate ao discurso de ódio por meio da moderação de conteúdo nas plataformas digitais, com foco na plataforma Youtube. Trata-se de pesquisa do tipo jurídico-interpretativo, cujo objetivo é analisar o estado da arte das discussões sobre a regulação do ambiente digital e a moderação de conteúdo relacionado ao discurso de ódio e identificar quais são os principais desafios para a efetivação da proteção a sujeitos vulneráveis contra esse tipo de violência *online* no Brasil. Para isso, a pesquisa se desenvolve em três capítulos principais. No primeiro capítulo, discute-se a regulação do ecossistema digital, considerando o papel da Internet na sociedade moderna e as implicações legais e sociais da sua regulação. O segundo capítulo foca no combate ao discurso de ódio, abordando as técnicas e desafios da moderação de conteúdo nas plataformas digitais, com destaque para a evolução e manifestações do discurso de ódio no ambiente digital. O último capítulo é um estudo de caso específico sobre o YouTube, examinando as políticas internas da plataforma para o combate ao discurso de ódio, os desafios enfrentados e as perspectivas futuras para melhorar a eficácia dessas políticas no Brasil. O trabalho argumenta que a teoria das vulnerabilidades pode oferecer uma contribuição significativa para a caracterização e regulação do discurso de ódio na Internet, ressaltando a necessidade de um equilíbrio entre liberdade de expressão e a proteção de grupos vulneráveis.

Palavras-chave: Moderação de conteúdo. Discurso de ódio. Regulação de plataformas. Vulnerabilidades digitais.

¹ Graduada em Direito pela Universidade Federal de Minas Gerais.

ABSTRACT

This study aims to investigate the issue of fighting hate speech through content moderation on digital platforms, with a focus on YouTube. This is a legal-interpretative study aimed at analyzing the current state of the art of discussions on the regulation of the digital environment and the moderation of content related to hate speech, as well as identifying the main challenges to ensuring the effective protection of vulnerable individuals against this form of online violence in Brazil. The research is developed across three main chapters. The first chapter discusses the regulation of the digital ecosystem, considering the role of the Internet in modern society and the legal and social implications of its regulation. The second chapter focuses on combating hate speech, addressing the techniques and challenges of content moderation on digital platforms, with an emphasis on the evolution and manifestations of hate speech in the digital environment. The final chapter is a specific case study on YouTube, examining the platform's internal policies for combating hate speech, the challenges faced, and future perspectives for improving the effectiveness of these policies in Brazil. The study argues that the theory of vulnerabilities can offer significant contributions to the characterization and regulation of hate speech on the Internet, highlighting the need for a balance between freedom of expression and the protection of vulnerable groups.

Keywords: Content moderation. Hate speech. Platform regulation. Digital vulnerabilities.

1 INTRODUÇÃO

*“– Então que máquina devo construir, mestre Medardo?
E o Bom começava a descrever-lhe a máquina que teria encomendado, e enriquecia a explicação traçando desenhos confusos. (...) E o carpinteiro era assaltado pela dúvida sobre se construir máquinas boas não estaria além das possibilidades humanas, ao passo que as únicas que de fato podiam funcionar com eficácia e exatidão seriam os patíbulos e as torturas”.
(O visconde partido ao meio. Calvino, 2009).*

A relação entre o desenvolvimento da tecnologia e sua interação com a sociedade é de reciprocidade e de redefinição mútua. Ao longo da história da ciência, o surgimento de novas técnicas frequentemente afeta o funcionamento e o ambiente social, e os usos empregados pelos indivíduos redefinem a utilização e o modo de se pensar a tecnologia. O questionamento sobre o controle dos impactos positivos e negativos da tecnologia na sociedade e sobre os novos dilemas e dificuldades que surgem com o desenvolvimento da ciência também não é uma particularidade moderna.

No conto “O visconde partido ao meio”, o escritor Calvino (2009) reflete sobre a questão de maneira bem-humorada ao imaginar a figura de um inventor que consegue cumprir com perfeição as instruções de um mestre maldoso para a criação de máquinas cada vez mais sofisticadas que têm o objetivo de infligir dor aos seres humanos. Um mestre bondoso, ao se

deparar com a evolução da técnica do inventor, decide instruí-lo a construir máquinas boas, que auxiliam no progresso da humanidade e no bem comum. Porém, as instruções se tornam cada vez mais complexas e inexecutáveis, e o autor, ao refletir sobre o descompasso do desenvolvimento da técnica com as necessidades da sociedade, pergunta ao leitor se “construir máquinas boas não estaria além das possibilidades humanas”².

Apesar do tom irônico e da natureza fictícia, a reflexão de Calvino (2009) representa uma preocupação comum, e cada vez mais evidente com a inserção de novas tecnologias da informação e comunicação no cotidiano social que alteram significativamente processos sociais e a própria forma de estruturação da sociedade. O historiador da tecnologia, Kranzberg (1986), ao refletir sobre o incipiente processo de mudança social pelo desenvolvimento de tecnologias digitais, teceu considerações que viriam a ser conhecidas como as “Leis de Kranzberg”, que sumarizavam sua visão sobre os processos de interação da tecnologia com a sociedade. Na primeira lei, Kranzberg (1986, p. 545) constata que “a tecnologia não é boa e nem má, mas tampouco é neutra”, e justifica:

Com isso, quero dizer que a interação da tecnologia com o ecossistema social ocorre de maneira na qual os desenvolvimentos técnicos frequentemente tem consequências ambientais, sociais e humanas que extrapolam os objetivos imediatos dos dispositivos e práticas, e que a mesma tecnologia pode ter resultados bastante diferentes quando introduzida em contextos diferentes ou sob circunstâncias distintas³.

A observação de Kranzberg (1986), embora elaborada no contexto de difusão e popularização da internet, marcado pelo otimismo com as possibilidades de liberdade e democratização do conhecimento na esfera digital, ainda se mostra relevante no contexto atual. O surgimento de desafios complexos relacionados ao mundo digital aponta para a necessidade de se analisar a relação da sociedade com a tecnologia, e formas de se controlar abusos e práticas nocivas que se disseminam com facilidade nesses espaços. Porém, a análise desses problemas contemporâneos perpassa pelos contextos e circunstâncias do uso da tecnologia, como apontado por Kranzberg (1986), e trazem o questionamento do inventor de

²CALVINO, Ítalo. **O visconde partido ao meio**. Tradução de Nilson Moulin. São Paulo: Companhia das Letras, 2009. p. 84.

³By that I mean that technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves, and the same technology can have quite different results when introduced into different contexts or under different circumstances (Kranzberg, 1986, p. 545-546).

KRANZBERG, Melvin. Technology and History: “Kranzberg's Laws”. **Technology and Culture**, Baltimore, v. 27, n. 3, p. 544–560, 1986. Disponível em: https://www.cs.ucdavis.edu/~koehl/Teaching/ECS188/PDF_files/Kranzberg.pdf. Acesso em: 15 jul. 2026. p. 05, tradução própria.

Calvino sobre a possibilidade de se conter riscos inerentes ao ambiente digital, amparado sobre uma estrutura de complexo desenvolvimento técnico, e de nele se promover finalidades comuns e interesses sociais.

Esta pesquisa se volta para a análise de um dos desdobramentos desse problema, que é a disseminação de conteúdo relacionado ao discurso de ódio no ambiente digital. Esse tipo de discurso não é um fenômeno contemporâneo, pois tradições jurídicas antigas já contemplavam dispositivos de proteção à honra pessoal e de vedação a ataques à personalidade em aspectos não físicos. Porém, as novas formas de circulação e de transmissão da informação que surgem com a inserção de tecnologias no cotidiano social, sobretudo plataformas digitais, tem como consequência o surgimento de novas manifestações do discurso de ódio com um grande potencial lesivo para grupos e comunidades vulneráveis. Elementos como a multiplicidade de atores na Internet, o contexto social no qual estão envolvidos, debates legislativos locais, interesses de grandes plataformas digitais e a conceitualização de elementos como o próprio discurso de ódio ou a vulnerabilidade formam um mosaico complexo que representa um enigma a ser decifrado por estudiosos e autoridades de forma a viabilizar a efetivação de princípios como a promoção da igualdade e a proteção da personalidade na era digital.

Diante desse cenário, esta pesquisa tem como objetivo analisar o atual estado da arte das discussões sobre a regulação do ambiente digital e a moderação de conteúdo relacionado ao discurso de ódio, e identificar, a partir das considerações teóricas, quais são os principais desafios para a efetivação da proteção a sujeitos vulneráveis contra esse tipo de violência *online* no Brasil.

A pesquisa classifica-se como do tipo jurídico-interpretativo ou jurídico compreensivo⁴, e utiliza as técnicas da análise de fontes de papel, como estratégia de pesquisa teórica⁵, e do estudo de caso, como estratégia de pesquisa empírica⁶. Adota-se como marco teórico o conceito de “ecossistema digital”, formulado por Almeida, Filgueiras e Doneda (2021), segundo o qual o ambiente digital – e, por consequência, a sua regulação – deve ser concebido de forma difusa e plural, a partir da compreensão de um espaço de interações complexas entre vários atores⁷.

⁴ GUSTIN, Miracy Barbosa de Souza; DIAS, Maria Tereza Fonseca; NICÁCIO, Camila Silva. **(Re)pensando a pesquisa jurídica: teoria e prática**. Belo Horizonte: Almeidina, 2020. p. 84.

⁵ GUSTIN, Miracy Barbosa de Souza; DIAS, Maria Tereza Fonseca; NICÁCIO, Camila Silva. **(Re)pensando a pesquisa jurídica: teoria e prática**. Belo Horizonte: Almeidina, 2020. p. 184.

⁶ GUSTIN, Miracy Barbosa de Souza; DIAS, Maria Tereza Fonseca; NICÁCIO, Camila Silva. **(Re)pensando a pesquisa jurídica: teoria e prática**. Belo Horizonte: Almeidina, 2020. p. 154.

⁷ ALMEIDA, Virgílio; FILGUEIRAS, Fernando; DONEDA, Danilo. The ecosystem of digital content

No primeiro capítulo, analisa-se o papel da Internet na sociedade moderna e a controvérsia relacionada à regulação do ecossistema digital. Apresenta-se algumas das abordagens apresentadas por pesquisadores tanto para a compreensão da interação das tecnologias digitais com a sociedade, quanto para propostas de regulamentação do ambiente digital.

No segundo capítulo, aborda-se a questão do combate ao discurso de ódio nas plataformas, a partir das considerações feitas acerca da regulação de ambientes digitais. Para isso, é feita uma análise sobre as principais técnicas e desafios para a moderação de conteúdo, e seu papel central no funcionamento das plataformas digitais. Em seguida, analisa-se a evolução do discurso de ódio até suas manifestações atuais no ambiente digital, propondo a compreensão deste fenômeno a partir da teoria das vulnerabilidades, diante do dever de efetivação da proteção do sujeito no ambiente digital e da consideração do contexto como elemento caracterizador de vulnerabilidades.

Por fim, no terceiro capítulo, realiza-se um estudo de caso sobre a caracterização e combate ao discurso de ódio na plataforma Youtube. Foram analisadas as políticas internas da plataforma e os relatórios de transparência disponibilizados pela empresa quanto políticas externas aplicáveis à questão no Brasil. Ao final, são apresentadas considerações sobre os desafios ao combate do discurso de ódio evidenciados pelo estudo de caso, diante do atual panorama regulatório da questão no Brasil.

Com este estudo, pretende-se apresentar uma contribuição da teoria das vulnerabilidades para a caracterização e regulação do discurso de ódio na Internet, e demonstrar, por meio da elaboração do estudo de caso, os principais empecilhos ao combate da disseminação da violência *online* no cenário atual brasileiro.

2 O ECOSSISTEMA DIGITAL: DILEMAS PARA A REGULAÇÃO DA INTERNET

2.1 Sociedade e Internet

O fenômeno da Internet redefiniu a forma de interação e de organização da sociedade moderna. Desde a sua criação, a Internet passou de uma ferramenta de comunicação restrita a pesquisadores e militares para uma verdadeira “teia mundial”, estruturada a partir de servidores e de protocolos de transmissão⁸.

Contudo, conforme Castells (2002), o ambiente digital não é caracterizado apenas por sua infraestrutura técnica. Mais do que uma rede de computadores interconectados, que viabilizam a comunicação e a colaboração em qualquer local e a qualquer tempo, pode-se argumentar que a Internet é um ecossistema complexo no qual elementos de natureza técnica e fenômenos sociais estão intrinsicamente ligados e se influenciam mutuamente. Assim, caracteriza-se como um fenômeno multifacetado que engloba aspectos técnicos, sociais, econômicos e culturais; e a sua compreensão deve levar em consideração a dualidade de sua natureza – como uma infraestrutura tecnológica complexa e como um espaço social dinâmico.

Nós sabemos que a tecnologia não determina a sociedade: ela é a sociedade. A sociedade é que dá forma à tecnologia de acordo com as necessidades, valores e interesses das pessoas que utilizam as tecnologias. Além disso, as tecnologias de comunicação e informação são particularmente sensíveis aos efeitos dos usos sociais da tecnologia⁹.

Inicialmente, a própria definição do que seria a Internet desperta divergências no campo científico. Destacam-se, de início, dois tipos de correntes: aquela que compreende a Internet como tecnologia e aquela que a define como um espaço social. Diante da multiplicidade de sentidos e definições atribuídas, falar sobre “a” Internet pode levar a um entendimento incorreto ou confuso, ou fornecer uma visão inadequada ou simplista da complexidade e diversidade inerentes ao conceito de Internet¹⁰.

A concepção da Internet como uma estrutura técnica e a sua definição em termos de

⁸ CASTELLS, Manuel. **A sociedade em rede**. Tradução de Roneide Venâncio Majer. São Paulo: Paz e Terra, 2002. v. 1. (A era da informação: economia, sociedade e cultura). p. 83-89.

⁹ CASTELLS, Manuel; CARDOSO, Gustavo. **A sociedade em rede: do conhecimento à acção política**. Belém, Portugal: Imprensa Nacional, Casa da Moeda, 2005. p. 16.

¹⁰ ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026. p. 14.

hardware e *software* que viabilizam a conexão e a transmissão de informação em escala global é a mais difundida no senso comum, mas também é este o posicionamento de parte dos estudiosos do tema e de atores relevantes no meio digital. O enfoque em componentes técnicos traz a vantagem da descrição da infraestrutura da internet para a finalidade de padronização, interoperabilidade e mecanismos de governança. Porém, esta representação encontra a limitação de “representar [a Internet] como um canal para a transmissão de dados, ao invés de um campo de prática social”¹¹, e desconsiderar a complexa interação entre a tecnologia e a sociedade.

Posicionamentos recentes que reconhecem a experiência da Internet como um espaço social, mais do que uma estrutura técnica, partem do pressuposto de que qualquer infraestrutura depende de ações humanas para a definição de sua padronização e de sua interação com outros sistemas. Ademais, a própria natureza de infraestrutura conferida a um sistema não é inerente à tecnologia: a integração da técnica com o cotidiano social faz com que ela seja moldada por finalidades comuns e convenções. A experiência da Internet como um espaço social é influenciada pela experiência local dos usuários – como a disponibilidade de máquinas e programas, provedores de serviço e aspectos culturais –, por aspectos políticos, como os que regulamentam a privacidade, e o ambiente social no qual o usuário está inserido¹². Tais particularidades locais da Internet, que a caracterizam como um espaço múltiplo, coexistem com forças e variáveis gerais ou supralocais que moldam a infraestrutura e o conteúdo que são disponibilizados, e com a existência de experiências similares e compartilhadas entre os usuários de todo o mundo¹³.

Castells (2002) propõe o conceito da “sociedade em rede” para a definição da nova forma de organização social que teria emergido da integração das novas tecnologias da comunicação e informação no cotidiano social. Para Castells (2013), o inevitável desenvolvimento técnico e científico, sobretudo em relação ao aumento da capacidade de processamento dos dispositivos eletrônicos, levaria a sociedade a um estágio no qual as redes

¹¹ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026. p. 3, tradução própria.

¹²ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026. p. 4 - 5.

¹³ ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026. ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026. p. 5.

se tornariam “a trama de nossa vida”¹⁴. Assim, a sociedade em rede seria definida por:

uma estrutura social baseada em redes operadas por tecnologias da informação e da comunicação fundamentadas na microeletrônica e em redes digitais de computadores que geram, processam e distribuem informação a partir de conhecimento acumulado nos nós dessas redes¹⁵.

No âmbito da reorganização da sociedade pela introdução de novas tecnologias, destaca-se, inclusive, que a inserção de tecnologias da informação no cotidiano social teria ocorrido de forma tão profunda que viabilizou a criação de novas formas de interação social que não poderiam ocorrer sem o suporte técnico da internet:

O ambiente virtual se tornou não somente locus de informação e comunicação, mas principalmente meio de trabalho, lazer, aprendizagem e socialização. Diz-se, por isso, que a atual “sociedade da interatividade” seria uma sociedade permeada por uma “camada digital” que suporta diversos tipos de atividades humanas em tese impossíveis sem a existência dessa tecnologia¹⁶.

Assim, a partir dessas considerações e para a finalidade do trabalho proposto, adota-se a definição de Internet como um espaço de característica sociotécnica. Dessa forma, contempla-se tanto a sua compreensão como uma complexa infraestrutura técnica, como a percepção de que a Internet também é um espaço social intrínseco ao cotidiano moderno. Este caráter dúplice também permite perceber a constante interação e redefinição mútua entre a tecnologia em si e a sociedade, que se evidencia na análise dos desafios enfrentados em um cotidiano social marcado pela interação praticamente indissociável com a tecnologia.

2.2 Regulação do ecossistema digital

Inicialmente, a popularização da Internet e a viabilização de novas formas de interação social, particularmente nas décadas de 1990 e 2000, geravam grande otimismo pelas possibilidades de democratização da informação, compartilhamento de ideias, e realização de debates em um espaço livre e gratuito. O ambiente digital e as plataformas eram vistas como um mecanismo para a emancipação, o esclarecimento e a realização de fins humanistas. O

¹⁴ CASTELLS, Manuel. **A sociedade em rede**. Tradução de Roneide Venâncio Majer. São Paulo: Paz e Terra, 2002. v. 1. (A era da informação: economia, sociedade e cultura). p. 91.

¹⁵ CASTELLS, Manuel; CARDOSO, Gustavo. **A sociedade em rede: do conhecimento à ação política**. Belém, Portugal: Imprensa Nacional, Casa da Moeda, 2005. p. 20.

¹⁶ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais: a contribuição possível da teoria da impolidez**. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 20.

professor e ativista Said (2007) considerava, por exemplo, que o ciberespaço representava “um campo democrático fantasticamente animador, aberto para todos os usuários de maneiras jamais sonhadas pelas gerações anteriores, tanto de tiranos quanto de ortodoxias”¹⁷, diante da constatação comum à época de que a Internet proporcionaria formas de comunicação e de colaboração que transcendem barreiras políticas, geográficas ou econômicas. A nova sociedade em rede representava acima de tudo possibilidades de subversão do controle político, social e cultural¹⁸.

A emergência de um novo espaço democrático fez com que alguns teóricos defendessem a ideia de que o ciberespaço não devesse sofrer interferências regulatórias externas como forma de proteção dos valores de liberdade, cooperação e colaboração que pautaram o *design* inicial da Internet. A rejeição de qualquer interferência estatal culminava em ideais como os expressos na “Declaração de Independência do Ciberespaço”¹⁹, segundo a qual o mundo virtual, espaço acessível a todos e sem privilégios ou preconceitos, deveria ser livre da influência de qualquer autoridade governamental.

Lessig (2006), ao retomar este momento histórico, sumariza que:

[...] a ideia dominante entre os entusiastas do ciberespaço era que este estava além do alcance da regulação do mundo físico. Os governos não poderiam interferir na vida online. E, assim, a vida online seria diferente, e separada, da dinâmica da vida online²⁰.

A Internet ainda é considerada como uma ferramenta que viabiliza grandes potenciais criativos e um “meio de interação síncrona e assíncrona, de promoção ou fomento de práticas de cidadania”²¹. Contudo, com a expansão do ciberespaço, que se tornou não só mais popular, mas também se converteu em um elemento indissociável de parte do cotidiano social e definidor de interações sociais nas mais diversas esferas, emergiram novos desafios e

¹⁷ SAID, Edward. **Orientalismo**: o Oriente como invenção do Ocidente. São Paulo: Companhia das Letras, 2007. p. 26.

¹⁸ LESSIG, Lawrence. **Code and other laws of cyberspace**: version 2.0. Nova Iorque, EUA: Basic Books, 2006. p. 05.

¹⁹ BARLOW, John Perry. **A Declaration of Independence of Cyberspace**. Suíça: Eletronic Frontier Foundation, 1996. Disponível em: <https://www.eff.org/pt-br/cyberspace-independence>. Acesso em: 10 jul. 2024.

²⁰ LESSIG, Lawrence. **Code and other laws of cyberspace**: version 2.0. Nova Iorque, EUA: Basic Books, 2006. p. ix, tradução própria.

“[...] the dominant idea among those who raved about cyberspace then was that cyberspace was beyond the reach of real-space regulation. Governments couldn’t touch life online. And hence, life online would be different, and separate, from the dynamic of life offline” (Lessig, 2006, p. ix).

²¹ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais**: a contribuição possível da teoria da impolidez. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 22.

conflitos. A utopia de uma Internet livre e democrática cede diante de desigualdades tanto ao acesso ao meio digital quanto de abusos e barreiras ocorridas no interior dele. O cenário tecnopolítico e tendências como “a concentração de mercado, a dificuldade de se discutir regulações, a opacidade das mediações algorítmicas e o avanço de práticas tóxicas e de intolerância”²² tornam evidente a necessidade de regulamentações e solução de disputas para as quais não há uma conclusão definitiva.

Para Almeida, Figueiras e Gaetani (2020) “a ênfase nos efeitos positivos e nas promessas de inovação e democratização do conhecimento deu lugar a uma preocupação crescente sobre ameaças nas redes e a um certo mal-estar em razão da hiperconectividade de quase tudo e de quase todos”²³.

Já Andrade (2019) destaca que a perda de parte de otimismo em relação à Internet como um espaço livre e autônomo tem como consequência a percepção da crescente necessidade de regulação em diversos aspectos:

A ausência de regulação se revelou inadequada quando se percebeu, enfim, que o ambiente virtual era tão suscetível ao abuso e à exploração do mais fraco pelo mais forte quanto o mundo real [...]. A multiplicação das ocorrências relacionadas a delitos virtuais, os conflitos de jurisdição em disputas comerciais e concorrenciais, as questões relacionadas à disparidade ou à segregação no acesso das pessoas a conhecimento, informação, bens e serviços disponíveis online, além de contendas especificamente relacionadas à segurança dos Estados e ao exercício de direitos políticos têm sido os principais fatores impulsionadores de uma crescente demanda internacional por regulação da internet²⁴.

Esses fatores dão ao ambiente digital características particulares que o diferem de outros espaços e sistemas sociais, no que diz respeito tanto ao seu funcionamento quanto à lógica a ser seguida para a sua regulação. Kleinwächter e Almeida (2015) identificam uma multiplicidade de regimes e camadas que coexistem dentro do ambiente digital, e propõe que a Internet seja compreendida como um “ecossistema complexo”, no qual “diversos atores com diferentes *status* legais operam em diversas camadas – em níveis locais, regionais, nacionais e

²² D’ANDREA, Carlos Frederico de Brito. **Pesquisando plataformas digitais: conceitos e métodos**. Salvador, Bahia: UFBA, 2020. p. 22.

²³ ALMEIDA, Virgílio; FILGUEIRAS, Fernando; DONEDA, Danilo. The ecosystem of digital content governance. **IEEE Internet Computing**, Washington, v. 25, n. 3, p. 13-17, 2021. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2021.3057756>. Acesso em: 15 jul. 2026. p. 43, tradução própria. “The emphasis on positive effects and on the promises of innovation and democratization of knowledge has given rise to a growing concern about threats on networks and a certain general malaise due to the hyperconnectivity of almost everything and everyone” (Almeida; Filgueiras; Doneda, 2021, p. 43).

²⁴ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais: a contribuição possível da teoria da impolidez**. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 23.

internacionais – motivados pela inovação tecnológica, necessidades do usuário, oportunidades de mercado e interesses políticos”²⁵.

Essa multiplicidade de camadas e regimes torna difícil a aplicação de uma regulação uniforme, uma vez que diferentes regiões e setores podem ter necessidades regulatórias distintas. Além disso, a própria natureza do ambiente digital apresenta características que dificultam a regulação tradicional. A Internet, por sua estrutura descentralizada e global, desafia as abordagens regulatórias tradicionais que se baseiam em jurisdições territoriais claras e normas legais específicas.

Além disso, como observado por Kleinwächter e Almeida (2015), dentro da Internet ocorre um “processo dinâmico” no qual uma vasta gama de diferentes regimes de regulação emergem, coexistem, se complementam ou se contrapõem, criando um ambiente em constante evolução. Dentre esses regimes, estariam tanto ambientes regulados pelo poder estatal quanto redes sem hierarquia, cuja estruturação se baseia em mecanismos de autorregulação. Os autores também destacam a existência de arranjos ou acordos para a correção de determinados espaços, nos quais há o influxo de interesses “de governos, do setor privado, da sociedade civil, e da comunidade técnica”²⁶. O dinamismo gera desafios significativos para legisladores e reguladores, que devem adaptar-se rapidamente às mudanças tecnológicas e aos novos modelos de negócio que surgem no ambiente digital²⁷.

Diante desses desafios, diversos modelos de regulação têm sido propostos e implementados para lidar com o ecossistema digital. Estudos na área de regulação de plataformas digitais indicam três principais abordagens que podem ser adotadas: a autorregulação, a correção e a heterorregulação²⁸. As abordagens se distinguem pelo grau de interferência estatal na organização e estruturação do ambiente digital:

O tripé comum a diversos modelos regulatórios autorregulação, heterorregulação e correção envolve leis e normas estabelecidas por entes distintos: o poder público, geralmente a partir de processo legislativo, no que se denomina

²⁵KLEINWÄCHTER, Wolfgang; ALMEIDA, Virgílio A.F. The internet governance ecosystem and the rainforest. **IEEE Internet Computing**, Washington, v. 19, n. 2, p. 64–67, 2015. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2015.49>. p. 64, tradução própria.

“[...] many players with different legal statuses operate on a variety of layers on local, national, regional, and international levels driven by technical innovation, user needs, market opportunities, and political interests” (Kleinwächter; Almeida, 2015, p. 64).

²⁶ KLEINWÄCHTER, Wolfgang; ALMEIDA, Virgílio A.F. The internet governance ecosystem and the rainforest. **IEEE Internet Computing**, Washington, v. 19, n. 2, p. 64–67, 2015. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2015.49>. p. 64, tradução própria.

²⁷ FILGUEIRAS, Fernando; ALMEIDA, Virgílio. **Governance for the digital world: neither more state nor more market**. Suíça: Palgrave Macmillan, 2021. p. 75, 97-104.

²⁸ RUEDIGGER, Marco Aurélio. **Regulação de plataformas digitais: uma regulação para a análise do debate nacional frente a um desafio global**. Rio de Janeiro: FGV DAPP, 2022. p. 09-12.

heterorregulação ou regulação pública; as empresas que oferecem o serviço, que podem estabelecer normas unilateralmente (auto-ordenação) ou definidas por uma associação de classe reconhecida pelo setor (autorregulação); ou uma interação entre ambos, no que se denomina corregulação ou autorregulação regulada, a depender do peso dado para um ou outro nessa combinação de forças²⁹.

A ausência de um consenso acerca da abordagem mais adequada evidencia-se ao comparar iniciativas e experiências internacionais de elaboração ou não de políticas públicas sobre o tema, e é indicador da complexidade e da persistente controvérsia da matéria. Valente (2019)³⁰, em um estudo comparativo sobre a regulação de plataformas digitais identifica que a autorregulação, que se caracteriza como a abstenção da responsabilidade de entes estatais, pode se manifestar pela absoluta ausência de políticas públicas ou pela adoção de medidas sem caráter vinculante pelas autoridades nacionais. Como exemplo, cita a experiência de países como o Canadá, os Estados Unidos e a Nigéria, nos quais há a tramitação de propostas legislativas sobre a regulação de aspectos do ambiente digital ou a elaboração de planos de educação midiática, mas cujo cenário ainda é marcado pela ausência de medidas vinculantes.

Quanto à corregulação, é mencionada a experiência da Comissão Europeia no âmbito da regulação das plataformas digitais. Optou-se pela elaboração de recomendações estruturais e vinculantes “ancorada em um código de práticas a ser adotado pelas plataformas digitais, com a formulação de planos de ação acerca de como tais diretrizes seriam implementadas”³¹. Assim, há a definição de princípios e práticas comuns por um documento vinculante, porém a sua implementação e efetiva concretização fica a cargo dos fornecedores de serviços digitais – no caso, as plataformas digitais.

Por fim, a abordagem da heterorregulação ou da regulação estatal representa a adoção de “medidas diretas de comando e controle”³² por parte da autoridade estatal em relação a determinado aspecto do ambiente digital. Valente (2019) menciona como exemplo a adoção de legislações específicas por países como a Malásia, Singapura e a Alemanha que regulam a circulação de informações falsas nas plataformas digitais, tipificando qual seria a conduta proibida, inclusive com a atribuição de pena de prisão em alguns casos. Também

²⁹ RUEDIGGER, Marco Aurélio. **Regulação de plataformas digitais**: uma regulação para a análise do debate nacional frente a um desafio global. Rio de Janeiro: FGV DAPP, 2022. p. 10.

³⁰ VALENTE, Jonas. Regulando desinformação e fake news: um panorama internacional das respostas ao problema. **Comunicação Pública**, Lisboa, v. 14, n. 27, p. 1-17, 2019. Disponível em: <https://journals.openedition.org/cp/5262#quotation>. Acesso em: 12 ago. 2024. p. 08-18.

³¹ VALENTE, Jonas. Regulando desinformação e fake news: um panorama internacional das respostas ao problema. **Comunicação Pública**, Lisboa, v. 14, n. 27, p. 1-17, 2019. Disponível em: <https://journals.openedition.org/cp/5262#quotation>. Acesso em: 12 ago. 2024. p. 10.

³² VALENTE, Jonas. Regulando desinformação e fake news: um panorama internacional das respostas ao problema. **Comunicação Pública**, Lisboa, v. 14, n. 27, p. 1-17, 2019. Disponível em: <https://journals.openedition.org/cp/5262#quotation>. Acesso em: 12 ago. 2024. p. 10.

pode haver a determinação expressa de adoção de mecanismos de transparência pré-determinados pelas plataformas, como no caso da Alemanha.

As abordagens regulatórias da Internet não são uniformes e dependem de elementos, como o objeto da regulamentação, o contexto social no qual a proposta está inserida, e elementos específicos do sistema digital que podem facilitar ou dificultar a implementação da proposta. Embora haja defesas de que uma abordagem descentralizada e cooperativa seria mais adequada à natureza difusa desse ambiente³³, a concentração de poderes nas mãos de atores com interesses frequentemente conflitantes, como Estados nacionais e grandes empresas, representam um obstáculo para a implementação desses ideais sobretudo em relação a tópicos sensíveis. Além disso, a natureza mutável do ecossistema digital e a necessidade de ratificação, seja voluntariamente ou por coerção, das políticas públicas por outros agentes, como provedores de serviço, para a sua efetiva implementação nos sistemas que se pretende regular³⁴ representam outros desafios para se pensar tanto a adequação de institutos já existentes para o mundo digital quanto a formulação de novas propostas.

3 COMBATE AO DISCURSO DE ÓDIO: MODERAÇÃO DE CONTEÚDO E PROTEÇÃO À PERSONALIDADE NO AMBIENTE DIGITAL

3.1 Moderação de conteúdo nas plataformas digitais

A moderação de conteúdo é uma prática encontrada em todas as formas de comunicação mediada. Uma forma de moderação de conteúdo é a seleção de informações e a adoção de escolhas editoriais em veículos de mídia tradicionais, como rádios, jornais e revistas. Estas formas de curadoria envolvem tanto a seleção do que será veiculado para o público quanto a definição de estratégias editoriais para atender a diversas finalidades³⁵.

No âmbito das plataformas digitais, também são identificadas formas de controle da informação veiculada³⁶, apesar da adoção de posicionamentos, por parte de algumas das

³³ ALMEIDA, Virgílio; FILGUEIRAS, Fernando; GAETANI, Francisco. Digital governance and the tragedy of the commons. **IEEE Internet Computing**, Washington, v. 24, n. 4, p. 41-46, 2020. Disponível em: <https://ieeexplore.ieee.org/document/9195909>. Acesso em: 15 jul. 2026.

³⁴ ALMEIDA, Virgílio; FILGUEIRAS, Fernando; DONEDA, Danilo. The ecosystem of digital content governance. **IEEE Internet Computing**, Washington, v. 25, n. 3, p. 13-17, 2021. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2021.3057756>. Acesso em: 15 jul. 2026. p. 83.

³⁵ KURTZ, Lahis Pasquali; CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. **Transparência da moderação de conteúdo: tendências regulatórias nacionais**. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. p. 9.

³⁶ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p.

empresas, de que a sua caracterização como “plataforma” representaria uma posição neutra de mera disponibilização de um espaço no qual usuários podem se conectar. Contudo, a curadoria, organização, arquivamento e moderação do conteúdo e da informação produzida pelos usuários é parte essencial da atividade das plataformas digitais: sem qualquer tipo de controle da informação produzida no espaço virtual, seria impossível que a plataforma oferecesse o serviço de determinada experiência *online* para os seus usuários.

Diante disso, algumas definições do próprio conceito de “plataforma digital” incluem a moderação de conteúdo como uma característica necessária. Gillespie (2018) afirma que a moderação de conteúdo não é uma simples atividade adicional das plataformas digitais, e sim um elemento “essencial, constituidor e definidor” de suas atividades. Para o autor (2018, p. 22), “não somente as plataformas não conseguem sobreviver sem a moderação de conteúdo, mas também não são plataformas sem ela”³⁷:

Plataformas digitais são: serviços e *sites online* que:

- a. hospedam, organizam e circulam o conteúdo compartilhado por usuários ou suas interações sociais,
- b. sem terem produzido ou encomendado o conteúdo (ou sua maior parte), são construídos sobre uma infraestrutura, abaixo da circulação da informação, c. que processa dados para oferecer serviços ao cliente, para publicidade, e para o lucro, [...]
- d. devem moderar o conteúdo e a atividade de seus usuários, utilizando mecanismos de detecção, revisão e execução de políticas³⁸.

Ao se considerar a moderação de conteúdo como uma realidade intrínseca ao funcionamento das plataformas digitais, afasta-se, como consequência, a posição de imparcialidade com a qual as provedoras pretendem se revestir. A seleção e curadoria de informação é uma parte do *design* do serviço e da experiência do usuário. Gillespie (2018) afirma que:

Compreender a moderação [de conteúdo] desta forma deve modificar nosso entendimento sobre o que as plataformas de mídia realmente fazem: de somente transmitir o que postamos, para constituir o que vemos. Não há posição de

939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 941.

³⁷ GILLESPIE, Talerton. **Custodians of the Internet: platforms, content moderation and the hidden decisions that shape social media**. New Haven: Yale University Press, 2018. p. 22, tradução própria.

³⁸ GILLESPIE, Talerton. **Custodians of the Internet: platforms, content moderation and the hidden decisions that shape social media**. New Haven: Yale University Press, 2018. p. 19, 22, tradução própria.

“For my purposes, platforms are: online sites and services that: host, organize, and circulate users shared content or social interactions for them, without having produced or commissioned (the bulk of) that content, built on an infrastructure, beneath that circulation of information, for processing data for customerservice, advertising, and profit, [...] platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement” (Gillespie, 2018, p. 19, 22).

imparcialidade. Os moderadores de plataforma escolhem e selecionam a todo momento, de diversas maneiras. Excluir pornografia ou ameaças, ou violência e terrorismo é apenas uma maneira pela qual as plataformas constituem o produto de mídia social que geram para a audiência. A crença persistente que plataformas são abertas, imparciais e irreguladas é bastante peculiar, considerando que *tudo* em uma plataforma é desenhado e orquestrado³⁹.

Embora a moderação de conteúdo seja uma realidade intrínseca ao funcionamento das plataformas digitais, ela pode ocorrer de diversas formas e se valer de mecanismos distintos. Ademais, formas de controle e de constituição de que informação os usuários veem e do conteúdo com o qual frequentemente se tornam ponto de debate público e político, diante da ampla popularidade das plataformas digitais. Na atualidade, as plataformas são campos importantes de formação de opinião, de disseminação de informação, e até mesmo de definição do debate político. Diante disso, o fato de a esfera pública digital estar “concentrada nas mãos de um pequeno número de plataformas digitais”⁴⁰ traz a preocupação de que as políticas de moderação de conteúdo priorizem interesses privados das empresas, visando o lucro e a promoção de filosofias internas, acima do interesse público e do bem comum.

A moderação de conteúdo é operacionalizada de maneira distinta por cada plataforma, em razão das suas particularidades de funcionamento, de infraestrutura, e de seus objetivos. Segundo Sander (2019), ela se efetivaria por “regras de governança privadas”, sejam elas implícitas, isto é, definidas pelos códigos e algoritmos que moldam as interações sociais na plataforma, ou explícitas, com o recurso a diversas técnicas de moderação⁴¹.

Embora existam inúmeras técnicas de moderação de conteúdo, algumas delas ganham notoriedade pelo seu uso difundido. Dentre as principais, destacam-se: remoção, indisponibilização, restrição, sinalização e ranqueamento⁴².

³⁹ GILLESPIE, Talerton. **Custodians of the Internet**: platforms, content moderation and the hidden decisions that shape social media. New Haven: Yale University Press, 2018. p. 22, tradução própria.

“Looking at moderation in this way should shift our view of what social media platforms really do: from transmitting what we post, to constituting what we see. There is no position of impartiality. Platform moderators pick and choose all the time, in all sorts of ways. Excluding porn or threats or violence or terrorism is just one way platforms constitute the social media product they are generating for the audience. The persistent belief that platforms are open, impartial, and unregulated is an odd one, considering that everything on a platform is designed and orchestrated” (Gillespie, 2018, p. 22).

⁴⁰ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 942.

⁴¹ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 946.

⁴² KURTZ, Lahis Pasquali; CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. **Transparência da moderação de conteúdo**: tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021.

Em primeiro lugar, a remoção de conteúdo consiste na eliminação completa de informações que violam as políticas da plataforma ou leis aplicáveis. Ao retirar de circulação conteúdos considerados inadmissíveis, a remoção visa minimizar os danos potenciais e proteger a comunidade de conteúdos prejudiciais. A eficácia da remoção depende da capacidade das plataformas de identificar rapidamente esses conteúdos, muitas vezes utilizando uma combinação de moderação automatizada e humana.

Já a indisponibilização envolve tornar um conteúdo inacessível para certos usuários ou regiões, sem removê-lo completamente da plataforma. Isso pode ocorrer temporariamente, enquanto uma revisão é realizada, ou permanentemente, se houver restrições legais que se apliquem a determinados locais geográficos. Por exemplo, um vídeo pode ser bloqueado em um país específico devido a leis locais de restrição de conteúdo, mas permanecer acessível em outras partes do mundo. Esse método permite que as plataformas cumpram regulações locais sem necessitar a exclusão global do conteúdo.

A restrição de conteúdo limita o acesso a determinados conteúdos com base em critérios específicos, como idade ou tipo de conteúdo. Plataformas como o YouTube utilizam restrições de idade para impedir que crianças e adolescentes acessem vídeos com temas violentos ou sexualmente explícitos ou que tratem de atividades perigosas, por exemplo⁴³. Além disso, conteúdos considerados sensíveis podem ser exibidos com avisos que informam os usuários sobre sua natureza, oferecendo a opção de continuar ou não a visualização. Como exemplo, o Instagram adota medidas de sinalização para conteúdo potencialmente gráfico ou sensível, limitando a visualização inicial do material com um aviso sobre a sua natureza. Somente caso o usuário consinta, a foto ou vídeo será exibida⁴⁴. Por fim, a restrição também pode incluir a desmonetização, que resulta na suspensão da geração de receita publicitária pelo conteúdo devido à natureza do material publicado. No Youtube, por exemplo, um dos requisitos para a monetização de vídeos e canais é a utilização exclusiva de material original do criador de conteúdo⁴⁵, como forma de coibir práticas de violação a direitos autorais.

Outra técnica, a sinalização, envolve anexar avisos ou etiquetas a conteúdos específicos para alertar os usuários para publicações patrocinadas, conteúdo potencialmente enganoso, impreciso ou equivocado, ou para publicações que abordem conteúdos

⁴³ YOUTUBE. **Age-restricted content**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2802167?hl=en>. Acesso em: 12 ago. 2024.

⁴⁴ INSTAGRAM. **Why you're seeing a warning before you can view a photo or video on Instagram**. Instagram, 2024. Disponível em: https://help.instagram.com/188848648282410?helpref=uf_permalink. Acesso em: 12 ago. 2024.

⁴⁵ YOUTUBE. **Youtube channel monetization policies**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/1311392?hl=en>. Acesso em: 12 ago. 2024.

controversos e sensíveis, como os de natureza eleitoral. A sinalização permite que os conteúdos permaneçam disponíveis, mas com contexto adicional que ajuda os usuários a fazerem julgamentos informados sobre o que estão vendo, e é uma técnica adotada pelas plataformas sobretudo no âmbito do combate à desinformação. O Facebook e o Instagram adotaram técnicas de sinalização ao implementarem um mecanismo que automaticamente detectava postagens que aludissem às eleições de 2022 no Brasil pelas palavras utilizadas, e exibia abaixo do conteúdo publicado um cartão com *links* para o site do Tribunal Superior Eleitoral (TSE)⁴⁶. O Twitter, atual X, também adotou uma estratégia de sinalização através do programa de “Notas da comunidade”, por meio do qual os próprios usuários podem adicionar observações exibidas em cartões junto com a postagem original, com o objetivo de informar sobre conteúdo controverso ou falso. Porém, a plataforma adota uma política de não moderar as notas da comunidade, e a definição de sua relevância fica a cargo dos próprios usuários. Isso leva muitas vezes à utilização da ferramenta para finalidades opostas como a disseminação de discurso de ódio por grupos organizados⁴⁷.

Por fim, o ranqueamento refere-se ao uso de algoritmos para determinar a visibilidade de conteúdos nas plataformas. Esses algoritmos avaliam uma variedade de fatores, como relevância, popularidade e comportamento do usuário, para priorizar quais conteúdos aparecerão no *feed* de notícias, resultados de busca ou recomendações. Essa técnica permite que plataformas destaquem conteúdos considerados mais pertinentes para os usuários, promovendo uma experiência de uso personalizada. Por outro lado, o ranqueamento também pode ser utilizado para ofuscar conteúdos considerados inadequados ou menos desejáveis, diminuindo sua visibilidade sem removê-los inteiramente.

As técnicas de moderação de conteúdo pode ser aplicadas antes ou depois da publicação⁴⁸. Um exemplo de moderação *ex ante* é a checagem de possíveis violações de direitos autorais ou de identificação de imagens conhecidas que contém alguma violação antes do *upload* definitivo pelo usuário. Já a moderação *ex post* revê o conteúdo que já está em

⁴⁶BEM PARANÁ. **Contra fake news, Instagram e Facebook colocam avisos em postagens sobre Eleições 2022.** Bem Paraná, Curitiba, 2021. Disponível em: <https://www.bemparana.com.br/publicacao/blogs/politicaemdebate/contra-fake-news-instagram-e-facebook-colocam-avisos-em-postagens-sobre-eleicoes-2022/>. Acesso em: 12 ago. 2024.

⁴⁷NALON, Tai. Notas da Comunidade do ex-Twitter são capturadas por grupos e expõem usuários ao ódio. **Terra**, São Paulo, 27 out. 2023. Disponível em: <https://www.terra.com.br/noticias/checamos/notas-da-comunidade-do-ex-twitter-sao-capturadas-por-grupos-e-expoem-usuarios-ao-odio,1874f0d37413940fe513015e62f3523as5luexff.html>. Acesso em: 21 ago. 2024.

⁴⁸SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 946.

circulação para detectar possíveis violações à política da plataforma e, caso seja detectada alguma desconformidade, encaminha a postagem a processos de revisão.

Ainda, a moderação de conteúdo pode se valer de diferentes abordagens. Sander (2019)⁴⁹ as classifica em três: a artesanal, comunitária e industrial. A abordagem artesanal realiza a moderação de conteúdo em pequena escala, com revisores humanos analisando publicações caso-a-caso. A moderação comunitária combina mecanismos da empresa, que estabelece regras e termos gerais da política da comunidade, com a contribuição de usuários voluntários ou “moderadores da comunidade” que fiscalizam a plataforma⁵⁰. Já a abordagem industrial se baseia em um sistema de larga escala para a análise do conteúdo publicado, que Sander (2019) classifica como “burocrático”, em razão da existência de regras elaboradas, da combinação de diversos processos – como o emprego de ferramentas automatizadas e da análise humana, e da divisão de tarefas em setores que analisam violações a políticas específicas da plataforma. Por exemplo, em abordagens industriais pode haver a criação de times de moderadores especializados em violação de direitos autorais, que trabalham separadamente de outros times voltados à análise de desinformação ou discurso de ódio.

Atualmente, especialmente nas grandes plataformas que se valem de uma abordagem industrial para a moderação de conteúdo, há o emprego de métodos que combinam a atuação humana e automatizada para a inspeção de conteúdo⁵¹. Algoritmos podem ser utilizados para analisar previamente determinado material, e tanto decidir que conteúdo será disponibilizado para qual usuário (ranqueamento), quanto inspecionar o material publicado para buscar padrões que podem representar violações à política da plataforma. Embora as técnicas automatizadas sejam indispensáveis para a moderação de conteúdo em larga escala, a decisão final de moderadores humanos ainda é essencial para o processo, pela necessidade de análise de contexto que muitas vezes surge para a caracterização ou não da violação a determinada política. Kurtz, Do Carmo e Vieira (2021) apresentam um fluxograma deste mecanismo de moderação de conteúdo que evidencia a combinação entre a atuação dos algoritmos e dos

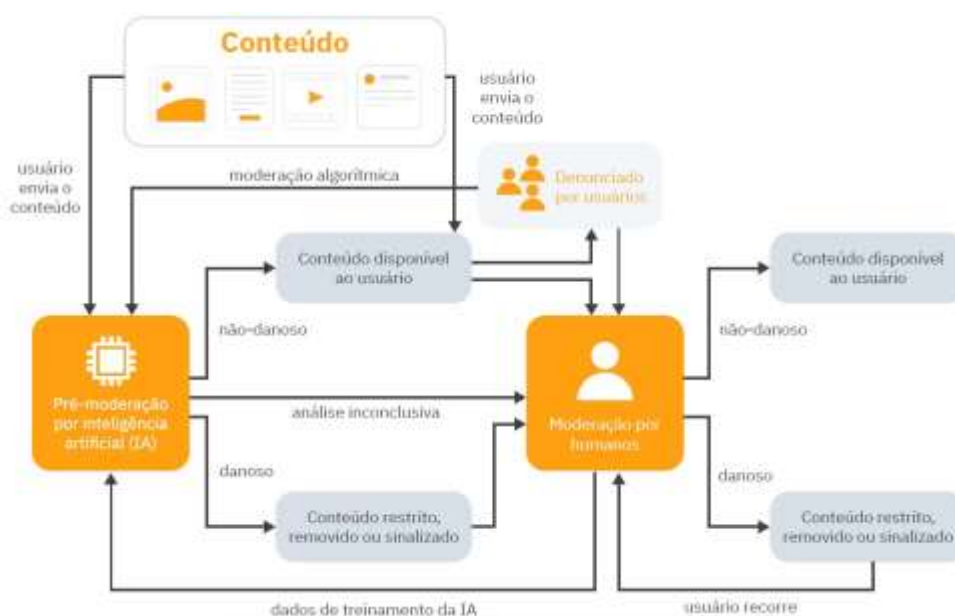
⁴⁹ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 947.

⁵⁰ Um exemplo de moderação comunitária é adotado pela plataforma *Reddit*, que é subdividida em fóruns voltados a diversas áreas de interesse. Os fóruns ou *subreddits* são criados e administrados pelos próprios usuários. Os membros da comunidade podem se tornar moderadores voluntários para aplicar as políticas da plataforma e regras específicas de convivência definidas em cada *subreddit*.

⁵¹ KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. **Transparência da moderação de conteúdo: tendências regulatórias nacionais**. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: <https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf>. Acesso em: 15 jul. 2026. p. 17.

moderadores humanos⁵²:

Figura 1 – Fluxo de moderação de conteúdo



Fonte: Kurtz, Do Carmo e Vieira (2021)

Por fim, embora a moderação de conteúdo seja operacionalizada pela própria empresa e se configure como um “mecanismo de governança privada”, as políticas e a forma de execução são influenciadas por fatores externos. Sander (2019)⁵³ identifica quatro principais fatores: a filosofia da empresa, a conformidade regulatória, a maximização de lucros e a opinião pública.

Inicialmente, a filosofia da empresa exerce uma influência significativa sobre as políticas de moderação de conteúdo. Cada plataforma digital tem uma missão e um conjunto de valores que orientam suas operações e decisões. Sander cita como exemplo o caso do Twitter, atual X, cuja política tradicional é a de supervalorização da liberdade de expressão dos seus usuários⁵⁴. A ausência de políticas de moderação provocou um ambiente de

⁵² KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. **Transparência da moderação de conteúdo: tendências regulatórias nacionais**. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: <https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf>. Acesso em: 15 jul. 2026. p. 16.

⁵³ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 948.

⁵⁴ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-

agressividade contra minorias e grupos vulneráveis e, conforme o autor, também levou à mudança gradual na filosofia da plataforma, que passou a instituir uma política mais restritiva de moderação de conteúdo⁵⁵ em reação a pressões externas. Portanto, as filosofias das empresas não apenas moldam as experiências dos usuários, mas também definem os limites do que é considerado aceitável na plataforma.

Já a conformidade regulatória é um fator externo que impõe um conjunto de obrigações legais às plataformas, que devem adaptar suas políticas para cumprir as leis vigentes em diversas jurisdições. Este fator é particularmente complexo para empresas globais que operam em múltiplos países com diferentes regimes legais. A necessidade de atender a regulamentos locais obriga as plataformas a implementarem medidas de moderação que podem variar significativamente de uma região para outra⁵⁶.

A maximização de lucros é um motivador central nas decisões de moderação de conteúdo. Plataformas digitais operam sob um modelo de negócios baseado na economia da atenção, onde o engajamento dos usuários é diretamente correlacionado à receita gerada por publicidade. Conteúdos que promovem alto engajamento, como notícias sensacionalistas ou polarizadoras, podem ser incentivados pelo algoritmo, mesmo que sejam controversos ou potencialmente danosos. Esse foco no lucro pode levar a decisões de moderação que priorizam a retenção do usuário sobre a qualidade e a segurança do conteúdo, resultando em ambientes online onde o conteúdo prejudicial não é adequadamente regulado. A pressão por lucros pode, portanto, comprometer a integridade das políticas de moderação e levantar questões sobre a responsabilidade das plataformas em proteger seus usuários contra danos⁵⁷.

Por fim, a opinião pública tem um impacto crescente sobre a moderação de conteúdo. Campanhas de ativismo digital, como o boicote #StopHateForProfit⁵⁸, realizado por marcas

rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 949.

⁵⁵ TWITTER/X. **Freedom of Speech, Not Reach**: an update on our enforcement philosophy. X Safety, Texas, 17 abr. 2023. Disponível em: https://blog.x.com/en_us/topics/product/2023/freedom-of-speechnot-reach-an-update-on-our-enforcement-philosophy. Acesso em: 12 ago. 2024.

⁵⁶ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 950.

⁵⁷ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 952.

⁵⁸ Em julho de 2020, a organização não-governamental *Anti-Defamation League* realizou uma campanha com a adesão de 1200 empresas para a suspensão da adesão dos negócios ao programa de anúncios do Facebook, que é fonte de renda para a empresa, para pressionar a plataforma a adotar medidas para combater a disseminação de ódio, racismo, antisemitismo e desinformação. A campanha foi nomeada de *Stop Hate for Profit* (ou “fim

investidoras em publicidade no Facebook, demonstram a tentativa de usuários e da sociedade civil para influenciar as políticas de conteúdo das plataformas⁵⁹. A resposta das empresas às críticas públicas pode levar a mudanças significativas nas políticas de moderação, como a implementação de novas regras contra discurso de ódio ou a maior transparência sobre processos de moderação, ou, em contrapartida, a ações de moderação que são mais reativas do que proativas. Ademais, embora essa forma de pressão seja relevante, pode haver lapsos temporais significativos entre a manifestação da opinião pública e a efetiva adoção de medidas pela plataforma⁶⁰.

3.2 Caracterização e manifestações do discurso de ódio na era digital

O discurso de ódio é um fenômeno que precede a Internet, e a sua proibição remonta a tradições jurídicas antigas. Proibições à difamação e a “condutas errôneas em relação a outrem [...] incluindo aspectos não-físicos da personalidade”⁶¹ estavam presentes em institutos do Direito Romano, e a vedação à instigação de ódio ou violência por meio da fala ou da conduta foi instituída em instrumentos legais, nacionais e internacionais, até a modernidade.

Porém, com a evolução de tecnologias da comunicação, que altera o modo de produção e difusão da informação, o discurso de ódio assume novas manifestações e alcance, com um grande potencial lesivo em relação a grupos vulneráveis que extrapola o ambiente digital. A supressão do discurso de ódio *online* é parte central do debate sobre a regulação do ambiente digital e da proteção da personalidade na *Internet*.

A definição proposta pelo jurista alemão Brugger, citado por Martins (2019), o caracteriza como o uso de “palavras que tendem a insultar, intimidar ou assediar pessoas em virtude de sua raça, cor, etnicidade, nacionalidade, sexo ou religião, entre outros atributos, tendo a capacidade de instigar violência, ódio ou discriminação contra tais pessoas”⁶², o que

ao ódio pelo lucro”), e parte da premissa que a priorização do lucro pelas plataformas ocorre em detrimento da proteção dos usuários e do combate a conteúdo nocivo.

⁵⁹ SLOANE, Garret. **Stop Hate for Profit**: os efeitos do boicote um ano depois. Meio e Mensagem, São Paulo, 2024. Disponível em: <https://www.meioemensagem.com.br/midia/stop-hate-for-profit-os-efeitos-do-boicote-um-ano-depois>. Acesso em: 8 ago. 2024.

⁶⁰ SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024. p. 954.

⁶¹ HIETANEN, Mika; EDDEBO, Johan. Towards a definition of hate speech with a focus on online contexts. Disponível em: **Journal of Communication Inquiry**, Londres, v. 47, n. 4, p. 441–458, out. 2022. <https://journals.sagepub.com/doi/epub/10.1177/01968599221124309>. Acesso em: 7 ago. 2024. p. 2.

⁶² MARTINS, Anna Clara Lehmann. Discurso de ódio em redes sociais e reconhecimento do outro: o caso M. **Revista Direito GV**, São Paulo, v. 15, n. 1, p. 1-30, jan./abr. 2019. Disponível em:

configura uma violência essencialmente simbólica, que pode ter efeitos tanto no plano do discurso quanto no plano físico. O discurso de ódio teria dois aspectos: o insulto, dirigido contra a pessoa ofendida com base em características de uma coletividade à qual pertence, e a instigação, que representa a conclamação à adesão à ideia discriminatória.

O discurso de ódio é dirigido contra uma coletividade e pode se efetivar pelo uso de estratégias como “a criação de estereótipos, a substituição de nomes, a seleção exclusiva de fatos favoráveis ao seu ponto de vista, a criação de inimigos, o apelo à autoridade, a afirmação e a repetição”⁶³. Porém, embora se encontrem manifestações frequentes baseadas em determinadas qualidades, como étnicas ou raciais, não é possível delimitar previamente que grupos podem ser alvo do discurso de ódio:

O conjunto dos grupos aos quais o discurso de ódio pode ser dirigido não é estático. Além dos já citados [características raciais, étnicos, nacionais, voltadas a sexo, religião, gênero e idade], atualmente é possível encontrar discurso de ódio contra certos círculos etários, de gênero e identidade de gênero, de origens regionais, de pessoas com deficiência entre outros. De maneira ampla, pode-se dizer que a norma protege coletividades que partilham uma determinada identidade sociocultural⁶⁴.

Ainda assim, a própria caracterização de o que é o discurso de ódio é objeto de disputas. As definições jurídicas e práticas desse fenômeno são por vezes consideradas “vagas ou contraditórias”, e a utilização de princípios como base para a sua descrição também resulta em ambiguidades. A dificuldade para a descrição do discurso de ódio advém de diversos fatores, dentre eles, as distintas formas de manifestação e percepção do que é o ódio, sobre as quais tampouco há consenso.

A Organização das Nações Unidas (ONU) divulgou, em 2019, um plano de ação para o discurso de ódio. O documento parte da premissa de que o crescimento da intolerância e do discurso de ódio nos meios de comunicação, sobretudo nos digitais, representa uma ameaça “aos valores democráticos, à estabilidade social e à paz”⁶⁵. A definição adotada para o discurso de ódio é:

<https://www.scielo.br/j/rdgv/a/WPZBfgrv6Md957dSxz7Hh5h/?format=pdf&lang=pt>. Acesso em: 7 nov. 2023. p. 60.

⁶³ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais**: a contribuição possível da teoria da impolidez. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 61.

⁶⁴ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais**: a contribuição possível da teoria da impolidez. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 61.

⁶⁵ ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **UN strategy and plan of action on hate speech**. ONU, Nova Iorque, Maio 2019. Disponível em: https://www.un.org/sites/un2.un.org/files/action_plan_on_hate_speech_en.pdf. Acesso em: 12 ago. 2024. p. 01.

[...] qualquer tipo de comunicação na fala, escrita ou comportamento, que ataca ou utiliza linguagem pejorativa ou discriminatória contra uma pessoa ou grupo com base em quem ela é, ou seja, com base em sua religião, etnia, nacionalidade, raça, cor, descendência, gênero ou outro fator de identidade⁶⁶.

Porém, a ONU reconhece que “não há uma definição legal no Direito Internacional do discurso de ódio”⁶⁷, e que a sua caracterização é controversa. Um dos fatores que contribuem para isso é a ausência de uma proibição expressa no direito internacional ao discurso de ódio em si. Os Estados-parte da ONU são obrigados a proibir manifestações do discurso de ódio que representam “incitação à discriminação, hostilidade e violência”, mas discute-se a extensão da obrigação dos Estados de também coibirem manifestações do discurso de ódio que não constituem incitação à violência.

Assim, no âmbito internacional dos direitos humanos, desenha-se um “limiar de gravidade” das manifestações do discurso de ódio. Em um relatório apresentado à Assembleia Geral no mesmo ano, o Comissariado para os Direitos Humanos reconhece que “certas expressões de ódio, por mais desagradáveis que sejam, não envolvem incitação ou ameaça direta, como declarações de preconceito contra grupos protegidos”.

A externalização desses sentimentos não seria diretamente proibida pelo conjunto de normas internacionais como o Pacto Internacional sobre Direitos Civis e Políticos e a Convenção Internacional para a Eliminação de Todas as Formas de Discriminação Racial. Portanto, o Comissariado recomenda que tais manifestações limítrofes do discurso de ódio sejam analisadas caso a caso, e que qualquer restrição do direito de livre expressão esteja amparada nos pilares da legalidade, necessidade e proporcionalidade⁶⁸:

A limitação ao discurso de ódio exige uma reconciliação de dois conjuntos de valores: os requisitos de uma sociedade democrática para a viabilização de um

⁶⁶ ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **UN strategy and plan of action on hate speech**. ONU, Nova Iorque, Maio 2019. Disponível em: https://www.un.org/sites/un2.un.org/files/action_plan_on_hate_speech_en.pdf. Acesso em: 12 ago. 2024. p. 2. “[...] any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor (ORGANIZAÇÃO DAS NAÇÕES UNIDAS, 2019, p. 2).

⁶⁷ ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **UN strategy and plan of action on hate speech**. ONU, Nova Iorque, maio 2019. Disponível em: https://www.un.org/sites/un2.un.org/files/action_plan_on_hate_speech_en.pdf. Acesso em: 12 ago. 2024. p. 2.

⁶⁸ ALTO COMISSÁRIO DAS NAÇÕES UNIDAS PARA OS DIREITOS HUMANOS. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/74/486)**. Nova Iorque: Assembleia Geral das Nações Unidas, out. 2019. Disponível em: <https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech>. Acesso em: 8 ago. 2024 p. 9-11.

debate aberto e da autonomia e desenvolvimento do indivíduo; e a obrigação de prevenir ataques contra comunidades vulneráveis e garantir a participação igualitária de todos os indivíduos na vida pública⁶⁹.

Organizações regionais e entidades nacionais também elaboram suas próprias definições de discurso de ódio para a realização de políticas e objetivos internos. No “Código de Conduta sobre discurso de ódio ilegal *online*” a Comissão Europeia define esse fenômeno como “a incitação pública à violência ou ao ódio com base em certas características, incluindo raça, cor, religião, descendência ou origem nacional ou étnica”⁷⁰.

A definição se aproxima daquela elaborada pela ONU, mas destaca-se a limitação do discurso de ódio apenas aos atos de incitação. Ademais, o próprio título do documento, que trata do “discurso de ódio ilegal”, remete à consideração feita pelo Comissariado de Direitos Humanos de que certas manifestações, como a externalização de opiniões e preconceitos, estariam fora do alcance da legislação de combate ao discurso de ódio, apesar de moralmente repreensíveis. O objetivo do código seria somente “remover discurso de ódio ilegal”, sob a justificativa de que “qualquer conteúdo considerado criminoso no mundo *offline* não deve ser autorizado na Internet”⁷¹. Outros conteúdos que “ofendam, choquem ou perturbem o Estado ou outra parcela da população” não são considerados ilegais, e tampouco há uma imposição da obrigação de retirada pela União Europeia⁷².

No sistema interamericano, não há um documento específico semelhante que defina e sistematize as medidas a serem adotadas pelos Estados para o combate ao discurso de ódio. Apesar disso, há instrumentos legais voltados para o combate à discriminação, como a “Convenção Interamericana para o combate a toda forma de discriminação e intolerância”,

⁶⁹ ALTO COMISSÁRIO DAS NAÇÕES UNIDAS PARA OS DIREITOS HUMANOS. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/74/486)**. Nova Iorque: Assembleia Geral das Nações Unidas, out. 2019. Disponível em: <https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech>. Acesso em: 8 ago. 2024. p. 4.

“[...] the limitation of hate speech seems to demand a reconciliation of two sets of values: democratic society's requirements to allow open debate and individual autonomy and development with the also compelling obligation to prevent attacks on vulnerable communities and ensure the equal and non-discriminatory participation of all individuals in public life” (ALTO COMISSÁRIO DAS NAÇÕES UNIDAS PARA OS DIREITOS HUMANOS, 2019, p. 4).

⁷⁰ JOROUVÁ, Vera. **Code of Conduct for Illegal online hate speech: questions and answers**. European Commission, Bruxelas, jun. 2016. Disponível em: https://ec.europa.eu/newsroom/document.cfm?doc_id=41844. Acesso em: 6 ago. 2024.

“Illegal Hate speech is defined in EU law as the public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin.”

⁷¹ JOROUVÁ, Vera. **Code of Conduct for Illegal online hate speech: questions and answers**. European Commission, Bruxelas, jun. 2016. Disponível em: https://ec.europa.eu/newsroom/document.cfm?doc_id=41844. Acesso em: 6 ago. 2024. p. 02.

⁷² JOROUVÁ, Vera. **Code of Conduct for Illegal online hate speech: questions and answers**. European Commission, Bruxelas, jun. 2016. Disponível em: https://ec.europa.eu/newsroom/document.cfm?doc_id=41844. Acesso em: 6 ago. 2024. p. 02.

que estabelece o dever dos Estados de prevenir, eliminar, proibir e punir “todos os atos e manifestações de discriminação e intolerância” (artigo 4º). É mencionada, expressamente, a circulação de materiais de comunicação nocivos:

Os Estados comprometem-se a prevenir, eliminar, proibir e punir, de acordo com suas normas constitucionais e com as disposições desta Convenção, todos os atos e manifestações de discriminação e intolerância, inclusive:

[...] ii. publicação, circulação ou difusão, por qualquer forma e/ou meio de comunicação, inclusive a internet, de qualquer material que:

- a) defenda, promova ou incite o ódio, a discriminação e a intolerância; e
- b) tolere, justifique ou defenda atos que constituam ou tenham constituído genocídio ou crimes contra a humanidade, conforme definidos pelo Direito Internacional, ou promova ou incite a prática desses atos⁷³.

Por fim, no ordenamento jurídico brasileiro também não há uma definição expressa do discurso de ódio. Contudo, as suas manifestações podem ferir bens jurídicos tutelados, o que confere embasamento legal para a sua coibição e combate. Um exemplo é o caso do direito à honra, que encontra amparo tanto na esfera criminal quanto na civil. No Código Penal, há a tipificação dos crimes de calúnia – a imputação falsa de fato definido como crime (artigo 138), difamação – a imputação de fato ofensivo (artigo 139) e injúria – ofensa à dignidade ou ao decoro de outrem (artigo 140). Já no Código Civil, a proteção à honra, à boa fama e à respeitabilidade é elemento dos direitos da personalidade (artigo 20), e assegura-se a indenização por injúria, calúnia e difamação (artigo 953), independentemente da prova de prejuízo material pelo ofendido.

A tipificação do crime de racismo, pela Lei 7.716/1989, também enumera atitudes discriminatórias proibidas, que consistem em formas de incitação ou promoção de violência com base em características identitárias de um grupo:

Art. 20. Praticar, induzir ou incitar a discriminação ou preconceito de raça, cor, etnia, religião ou procedência nacional. (...)

§1º Fabricar, comercializar, distribuir ou veicular símbolos, emblemas, ornamentos, distintivos ou propaganda que utilizem a cruz suástica ou gamada, para fins de divulgação do nazismo. (...)

§2º Se qualquer dos crimes previstos neste artigo for cometido por intermédio dos meios de comunicação social, de publicação em redes sociais, da rede mundial de computadores ou de publicação de qualquer natureza (...)

§2º-A Se qualquer dos crimes previstos neste artigo for cometido no contexto de atividades esportivas, religiosas, artísticas ou culturais destinadas ao público: (...)

⁷³ ORGANIZAÇÃO DOS ESTADOS AMERICANOS. Convenção Interamericana para o combate a toda forma de discriminação e intolerância. OEA, Guatemala, 05 jun. 2013. Disponível em: https://www.oas.org/en/sla/dil/docs/inter_american_treaties_A-69_Convencao_Interamericana_disciminacao_intolerancia_POR.pdf. Acesso em: 8 ago. 2024.

§2º-B Sem prejuízo da pena correspondente à violência, incorre nas mesmas penas previstas no caput deste artigo quem obstar, impedir ou empregar violência contra quaisquer manifestações ou práticas religiosas.

A Lei nº 7.716/1989 não possui o objetivo de combater o discurso de ódio, e sim os crimes resultantes de preconceito de raça, cor, etnia ou religião, os quais podem se manifestar por meio do discurso de ódio. Assim, no cenário jurídico nacional, percebe-se a mesma situação problematizada pela ONU: certas manifestações mais extremas do discurso de ódio são vedadas por legislações que tutelam outros bens jurídicos. Porém, questiona-se qual é o tratamento a ser dado a outras formas de expressão de ódio, preconceito e discriminação com base na identidade de grupo, que, embora manifestadas de forma menos extrema, podem ser danosas aos grupos vulneráveis. Expressões mais leves do discurso de ódio são frequentes nas redes sociais⁷⁴, e a ausência de uma definição transparente e que possa ser efetivamente operacionalizada na moderação de conteúdo representa um primeiro obstáculo ao combate ao discurso de ódio, sobretudo na Internet⁷⁵.

As manifestações do discurso de ódio *online* são diversas, e as plataformas digitais frequentemente enfrentam críticas pelo tratamento inadequado do problema. A título exemplificativo, citam-se duas situações em que o discurso de ódio nas redes sociais foi levado à apreciação do Poder Judiciário.

O primeiro caso se trata de Ação Civil Pública proposta pelo Ministério Público Federal (MPF) em 2021⁷⁶ contra um apresentador de televisão pela divulgação de um vídeo publicado no Youtube em 2018, que contém falas discriminatórias contra o gênero feminino e ataques a uma mulher identificada na gravação, que à época se encontrava sob custódia do Estado da Paraíba. No trecho do programa, em coro com outros participantes da produção, o apresentador replica falas como:

Pau que nasce torto... nunca se endireita! Menina que requebra... pega na cabeça!
Mulher que não pinta a unha é sebosa! (...) Mulher que não passa a lâmina no, no, na

⁷⁴ ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais**: a contribuição possível da teoria da impolidez. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023. p. 62.

⁷⁵ HIETANEN, Mika; EDDEBO, Johan. Towards a definition of hate speech with a focus on online contexts. Disponível em: **Journal of Communication Inquiry**, Londres, v. 47, n. 4, p. 441–458, out. 2022. <https://journals.sagepub.com/doi/epub/10.1177/01968599221124309>. Acesso em: 7 ago. 2024. p. 441.

⁷⁶ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Inquérito Civil n. 1.24.000.000812/2021-0**. Paraíba, out. 2019. Disponível em: <https://uploads.polemicaparaiba.com.br/2021/06/ACP-Discurso-de-odio-contra-mulheres-na-Paraiba.pdf>. Acesso em: 14 ago. 2024.

[o apresentador faz gestos de depilação de genitálias e de ambas as axilas], ela é o quê? É sebosa! É sebosa! [sic]⁷⁷.

Em seguida, é projetada a imagem da mulher em uma tela, e a ela o apresentador também dirige comentários pejorativos e misóginos:

Respeite sua família, sua vagabunda preguiçosa! Cadê o que está escrito na camisa dela? Ah, “eu não sou princesa. Eu sou a rainha do pop star.” [...] Ela é uma pop star. Uma pop star do fumo! Pop star pó! Né, pó estar, aqui nessa venta! Venta de jumenta! Isso deve cheirar que é uma beleza. Botar ela para cheirar na cadeia era água sanitária. Com um rodo bom para ela lavar as salas, aprender a lavar e ser dona de casa. E outra, ela não pinta as unhas! [sic]⁷⁸

Como fundamentação, o MPF alegou que o conteúdo nocivo do vídeo afeta tanto os direitos fundamentais da mulher identificada quanto a coletividade de mulheres como um todo, em razão do tom pejorativo empregado. Na petição, as “falas discriminatórias” são equiparadas ao discurso de ódio. Contudo, ao se analisar o conteúdo das falas, identifica-se tanto o elemento do insulto, ou seja, o ataque à vítima com base na coletividade à qual pertence (o gênero feminino), quanto a conclamação à adesão à ideia discriminatória, ou instigação, por meio da criação de estereótipos e da repetição das falas pelos participantes do programa.

Até a data de encerramento desta pesquisa⁷⁹, o vídeo ainda estava no ar, com 8.3 mil curtidas. Nos comentários, há usuários que criticam a conduta do apresentador, mas também comentários de apoio ao discurso propagado pelo vídeo e críticas à atuação do Ministério Público e do Poder Judiciário após a repercussão do caso. Em 2023, o MPF ofereceu denúncia contra o apresentador pela prática do crime de preconceito racial de gênero na mesma ocasião⁸⁰. Porém, ainda não foi prolatada sentença em nenhuma das duas ações até o fechamento da pesquisa⁸¹.

O segundo caso se trata de Ação Civil Pública ajuizada pelo MPF e pela Defensoria Pública da União (DPU) em face da Google, do Estado do Rio de Janeiro e de policiais

⁷⁷ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Inquérito Civil n. 1.24.000.000812/2021-0.** Paraíba, out. 2019. Disponível em: <https://uploads.polemicaparaiba.com.br/2021/06/ACP-Discorso-de-odio-contra-mulheres-na-Paraiba.pdf>. Acesso em: 14 ago. 2024. p. 5.

⁷⁸ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Inquérito Civil n. 1.24.000.000812/2021-0.** Paraíba, out. 2019. Disponível em: <https://uploads.polemicaparaiba.com.br/2021/06/ACP-Discorso-de-odio-contra-mulheres-na-Paraiba.pdf>. Acesso em: 14 ago. 2024. p. 4.

⁷⁹ Agosto de 2024.

⁸⁰ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Procedimento Investigativo Criminal n. 1.24.000.000924/2021-51.** Paraíba, jan. 2023. Disponível em: <https://static.poder360.com.br/2023/02/pedido-prisao-sikera-junior-mpf.pdf>. Acesso em: 14 ago. 2024.

⁸¹ Agosto de 2024

militares que veiculavam conteúdo violento e de discurso de ódio voltado contra populações periféricas em canais do Youtube⁸². Identificou-se que determinados canais administrados por policiais militares, em formato de *podcast* ou *videocast*, continham vídeos nos quais os agentes descrevem abordagens violentas, confrontos, e agressões a moradores de comunidades ou a recrutas da própria Polícia. Os vídeos possuiriam tom cômico, descrições pejorativas da população afetada pelas operações, apologia e incitação a comportamentos violentos. As entrevistas listadas pelo MPF e pela DPU na petição inicial foram retiradas do ar⁸³, mas possuíam títulos como “Fiquei sem controle no BOPE” e “Cachorro louco conta como quebrou 4 em favela”, e conteúdo extremamente violento e depreciativo:

Aí o maluco tipo assim viu a cena, cresceu. Nisso que ele cresceu, aí eu descontrolei. Aí foi agressividade descontrolada. O maluco veio pra cima, eu peguei o maluco e já comecei ‘pau, pau, porrada’ [gesto de soco], o maluco tomando, caindo no chão, eu segurando a cara dele, dando, aí eu já fiquei cego. Aí eu sei que veio uma mulher, veio uma grávida, veio mais gente, eu peguei a mulher e joga a mulher pra lá, e a grávida foi pra lá (...) e a grávida voltou, eu enrolei ela pelo cabelo e tomou (gesto de tapa), voou longe [sic]⁸⁴.

Em outros vídeos, são descritas operações policiais, como a tomada do morro do Alemão e da Penha em 2010:

Apresentador: E prendeu pouco e matou pouco, né?
Sargento: Cara, [no morro do Alemão] não teve um tiro, não teve um tiro, mas na Penha teve bala pra c***! [...]
Apresentador: O que faltou, irmão, o que faltou ali naquela cena? De autorização?
Sargento: Cara, eu acho que se botasse uma equipe na mata resolvia bastante coisa. Pegasse uma equipe do BOPE e colocasse na mata, acho que a gente conseguia (...), a gente ia mostrar pro tráfico pra ele não fazer mais isso. “Ah, mas matou pra c***!”
Tinha que ser, pô, entendeu? Tá tacando fogo, acho que uma coroa morreu incendiada num ônibus. Pô, imagina? Cara, ridículo, cara... entendeu? Cara, isso é terrorismo. Qualquer país civilizado, o que eles fazem é terrorismo. Só aqui no Brasil que a p*** de maconheiro que gosta de ficar... é brincar de espada, que acha que os caras são santos, é complicado, entendeu? [sic]⁸⁵.

Além da propagação de ofensas e insultos a indivíduos específicos e à divulgação

⁸² MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República no Rio de Janeiro. **Ação Civil Pública n. 5028178-59.2024.4.02.5101**. Rio de Janeiro, abr. 2024. Disponível em: <https://tinyurl.com/4ddzy6ec>. Acesso em: 14 ago. 2024.

⁸³ RIO DE JANEIRO, TRF2. **Ação Civil Pública n. 5028178-59.2024.4.02.5101**. Abr. 2024. Disponível em: <https://tinyurl.com/4ddzy6ec>. Acesso em: 14 ago. 2024, p. 04-05.

⁸⁴ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República no Rio de Janeiro. **Ação Civil Pública n. 5028178-59.2024.4.02.5101**. Rio de Janeiro, abr. 2024. Disponível em: <https://tinyurl.com/4ddzy6ec>. Acesso em: 14 ago. 2024, p. 6.

⁸⁵ MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República no Rio de Janeiro. **Ação Civil Pública n. 5028178-59.2024.4.02.5101**. Rio de Janeiro, abr. 2024. Disponível em: <https://tinyurl.com/4ddzy6ec>. Acesso em: 14 ago. 2024, p. 12.

pública de atos das autoridades que violam prerrogativas do Estado de Direito, identifica-se na descrição dos vídeos atos de seleção de fatos favoráveis ao ponto de vista dos policiais envolvidos, de criação de inimigos e de apelo à autoridade, o que permite caracterizar o conteúdo como discurso de ódio segundo a definição de Brugger.

Neste caso, o MPF e a DPU requereram ao final a derrubada do conteúdo e a responsabilização da Google/Youtube e dos administradores dos canais a pagamento de indenização, tal como a regulamentação do discurso de ódio para policiais militares pela Secretaria de Estado do Rio de Janeiro⁸⁶. Até o encerramento da pesquisa⁸⁷, não havia sido prolatada sentença no caso, apesar de ter sido deferida parcialmente a tutela de urgência para a suspensão do acesso aos vídeos listados, sem sua exclusão definitiva.

A disseminação do discurso de ódio na Internet pode também causar impactos concretos fora do mundo virtual. A apuração de episódios recentes de ataques violentos e crimes de ódio aponta a participação em comunidades voltadas para a propagação de discurso de ódio digital como uma das motivações para os atos de violência direta. Um exemplo é o ataque à sinagoga de Pittsburgh, nos Estados Unidos, em 2018⁸⁸, no qual se identificou que o atirador havia postado mensagens antissemitas nas redes sociais antes de cometer o crime. Da mesma forma, o ataque realizado em Christchurch, Nova Zelândia, em 2019⁸⁹, que teve como alvo muçulmanos que frequentavam uma mesquita da cidade, foi perpetrado por um indivíduo que havia sido influenciado por comunidades online de extrema direita, nas quais havia a frequente disseminação de discurso de ódio contra imigrantes e muçulmanos.

Embora o discurso de ódio tenha consequências graves no mundo *offline*, também possui efeitos nocivos quando se limita à própria esfera virtual. Argumenta-se que não é necessário que o discurso de ódio produza efeitos fora da internet para que esteja caracterizada a violência que deve ser combatida. As plataformas digitais representam espaços cada vez mais relevantes de convivência social, de expressão do sujeito e de sua personalidade, e de participação no debate público. Assim, a incitação a comportamentos violentos e à discriminação e a prática de insultos repetidos no discurso de ódio também

⁸⁶ CONSULTOR JURÍDICO. MPF e DPU acionam Google na Justiça por discurso de ódio de policiais. **Consultor Jurídico**, São Paulo, 4 maio 2024. Disponível em: <https://www.conjur.com.br/2024-mai-04/mpf-e-dpu-acionam-google-na-justica-por-discurso-de-odio-de-policiais/>. Acesso em: 14 ago. 2024.

⁸⁷ Agosto de 2024.

⁸⁸ CHAVEZ, Nicole et al. **Pittsburgh synagogue gunman said he wanted all Jews to die, criminal complaint says**. CNN, Estados Unidos, 2018. Disponível em: <https://edition.cnn.com/2018/10/28/us/pittsburgh-synagogue-shooting/index.html>. Acesso em: 6 ago. 2024.

⁸⁹ MACKLIN, Graham. The Christchurch Attacks: livestream terror in the viral video age. **Combating Terrorism Center**, Nova Iorque, v. 12, n. 6, jul. 2019. Disponível em: <https://ctc.westpoint.edu/christchurch-attacks-livestream-terror-viral-video-age/>. Acesso em: 8 ago. 2024.

configuram, por si sós, a prática de violência *online*. Acerca disso, Poland (2016), ao analisar os efeitos da violência digital contra mulheres, o que intitula de “cybersexismo”, faz a seguinte consideração:

Um dos métodos mais comuns para minimizar os efeitos da violência *online* é insistir que a Internet não é um espaço que deve ser levado a sério. Nessa abordagem, a Internet é apresentada como um espaço irreal ou liminar que não pode ter efeitos *offline*, porque a agressão na Internet não seria tão “real” quando a agressão *offline*. [...] Hoje, porém, a Internet é tão real e importante para a maioria das pessoas como a vida *offline*. De países em desenvolvimento às nações industrializadas, todos estão conectados, e o uso da Internet continua a crescer. A fronteira entre a vida *online* e *offline* se torna mais tênue, o que significa que a linha entre a agressão *online* e *offline* também se apaga – e os efeitos de ambos os tipos de violência são incrivelmente parecidos⁹⁰.

Por fim, outra característica das manifestações do discurso de ódio nas plataformas digitais é que estas frequentemente são subjetivas e dependem do contexto para serem identificadas. Isso dificulta processos de moderação de conteúdo automáticos baseados na mera análise textual do material publicado. Em um estudo sobre a caracterização e detecção de usuários propagadores de discurso de ódio no Twitter (atual X), pesquisadores identificaram que a incitação de violência com base na identidade de grupo por vezes ocorria em postagens sem apologia a palavras ou símbolos reconhecidamente violentos, e a sua caracterização seria possível apenas pela análise de contexto. Por exemplo, foi identificado o *tweet*: *Timesup, yall getting w should have happened long ago* (“finalmente, está acontecendo com vocês o que deveria ter acontecido há muito tempo”)⁹¹ em resposta a uma postagem que mencionava o Holocausto. Embora esta seja claramente uma manifestação do discurso de ódio, é impossível detectá-la pela análise isolada do texto, e o volume de conteúdo a ser checado também impossibilita a análise manual de todas as postagens por moderadores humanos⁹².

Diante disso, com base na estrutura da plataforma, os autores sugerem um modelo de

⁹⁰ POLAND, Baley. **Haters: harassment, abuse and violence online**. Nebraska, EUA: Potomac Books, 2016. p. 90, tradução própria.

“One of the most common methods for dismissing the effects of online abuse is to insist that the Internet is not a space that we should take seriously. In this approach the Internet is presented as an unreal or liminal environment that cannot have offline effects, because Internet harassment is not as real as offline harassment. [...] Today, however, the Internet is as real and as important for most people as offline life. From developing countries to industrialized nations, practically everyone is connected, and rates of Internet use continue to increase. The lines between online and offline life continue to blur, which means the lines between online and offline harassment also continue to blur and the effects of both types of abuse are incredibly similar” (Poland, 2016, p. 90).

⁹¹ RIBEIRO, Manoel Horta et al. **Characterizing and Detecting Hateful Users on Twitter**. 23 mar. 2018. Disponível em: <https://arxiv.org/pdf/1803.08977>. Acesso em: 8 ago. 2024. p. 1.

⁹² RIBEIRO, Manoel Horta et al. **Characterizing and Detecting Hateful Users on Twitter**. 23 mar. 2018. Disponível em: <https://arxiv.org/pdf/1803.08977>. Acesso em: 8 ago. 2024. p. 1.

moderação de conteúdo baseado na caracterização dos usuários propagadores do discurso de ódio, e não na análise textual das postagens. Determinadas postagens podem parecer inofensivas fora de contexto, mas, quando analisadas em conjunto com o perfil do usuário e suas conexões, revelam um padrão de comportamento odioso. A análise do comportamento destes usuários revela que adotam padrões de uso da rede social distintos dos usuários comuns, como: a criação mais recente de contas; uso frequente de frases negativas; uso menos frequente de vocabulário relacionado ao ódio, terrorismo e violência e mais frequente de palavras ligadas ao amor, trabalho e masculinidade; e uma conexão mais densa em redes de compartilhamento⁹³. Assim, abordagens de detecção baseadas em grafos, que consideram as conexões entre os usuários, superam métodos tradicionais de análise de texto isolado⁹⁴.

A proposta de detecção automática do discurso de ódio com base no contexto e no comportamento de usuários também já foi objeto de outros estudos⁹⁵. Contudo, este enfoque para a moderação de conteúdo depende da análise da estrutura e do funcionamento de cada plataforma para ser adequadamente implementado, já que a análise do comportamento e das formas de conexão sociais entre os usuários propagadores de discurso de ódio varia significativamente em razão das dinâmicas das diferentes redes sociais.

3.3 Vulnerabilidades digitais e o discurso de ódio

A vulnerabilidade, no contexto jurídico, é um conceito central que remete à ideia de fragilidade e à necessidade de proteção dos indivíduos em situações de desvantagem. A vulnerabilidade também justifica a proteção diferenciada de certos grupos ou indivíduos que, por suas características ou circunstâncias, estão em posição de desvantagem em relação a outros atores, como empresas ou grandes organizações.

Conforme Marques e Mucelin (2022)⁹⁶, a vulnerabilidade possui um duplo aspecto: por um lado, ela aponta para a fragilidade inerente a certas situações ou condições; por outro, ela sublinha a necessidade de proteção que emerge dessa fragilidade. No contexto das

⁹³ RIBEIRO, Manoel Horta et al. **Characterizing and Detecting Hateful Users on Twitter**. 23 mar. 2018. Disponível em: <https://arxiv.org/pdf/1803.08977>. Acesso em: 8 ago. 2024. p. 09.

⁹⁴ RIBEIRO, Manoel Horta et al. **Characterizing and Detecting Hateful Users on Twitter**. 23 mar. 2018. Disponível em: <https://arxiv.org/pdf/1803.08977>. Acesso em: 8 ago. 2024. p. 03.

⁹⁵ GAO, Lei; HUANG, Ruihong. **Detecting Online Hate Speech Using Context Aware Models**. 2018. Disponível em: <https://arxiv.org/abs/1710.07395>. Acesso em: 13 ago. 2024.

⁹⁶ MARQUES, Claudia Lima; MUCELIN, Guilherme. Vulnerabilidade na era digital: um estudo sobre os fatores de vulnerabilidade da pessoa natural nas plataformas, a partir da dogmática do Direito do Consumidor. **Civilistica.com**, Rio de Janeiro, v. 11, n. 3, p. 1-30, dez. 2022. Disponível em: <https://civilistica.emnuvens.com.br/redc/article/view/872>. Acesso em: 7 nov. 2023. p. 02.

relações privadas, essa vulnerabilidade pode manifestar-se de diversas formas, seja no desequilíbrio de poder entre consumidores e fornecedores, seja na assimetria de informações, onde o consumidor frequentemente não possui o mesmo nível de conhecimento técnico ou jurídico que o fornecedor.

No Brasil, o interesse do direito pela vulnerabilidade ganhou destaque com a edição do Código de Defesa do Consumidor (CDC) em 1990, que formalizou o reconhecimento da vulnerabilidade do consumidor como princípio fundamental, inicialmente por razões de cunho essencialmente econômico. Contudo, a ideia de vulnerabilidade como fundamento para a proteção existencial, inclusive no âmbito de relações privadas, se estende para outros grupos sociais com o desenvolvimento do direito brasileiro após a Constituição da República de 1988 (CR/88), como idosos, crianças e pessoas com deficiência, refletindo uma preocupação crescente com a intervenção nas relações sociais como forma de reequilibrar desigualdades e proteger a dignidade da pessoa humana⁹⁷.

Luna (2009) propõe que o conceito de vulnerabilidade seja compreendido de maneira dinâmica, e não estática, e desenvolve o conceito das capas ou camadas de vulnerabilidade. Ao invés de tratar a vulnerabilidade como uma condição fixa, que se operaria como rótulos que são atribuídos a uma pessoa que é ou não vulnerável, a metáfora das camadas sugere que a vulnerabilidade é algo maleável, composto por várias dimensões que podem variar de acordo com o contexto e as circunstâncias⁹⁸. Assim, não existiria um conjunto de vulnerabilidades fixo e pré-determinado, e a caracterização da situação de vulnerabilidade estaria inerentemente relacionada ao contexto social no qual o sujeito se encontra. A vulnerabilidade, portanto, não é um estado estático, mas uma condição dinâmica que varia conforme o contexto social, econômico e tecnológico em que o sujeito está inserido:

A metáfora da camada dá a ideia de algo mais maleável, algo que pode ser múltiplo e diferente, e que pode ser removido camada por camada. Não há uma vulnerabilidade sólida e única que exaure a categoria; pode haver diferentes vulnerabilidades, diferentes camadas operando. Essas camadas podem se sobrepor: algumas podem [...] estar relacionadas a circunstâncias sociais. A ideia de camadas de vulnerabilidade dá flexibilidade a este conceito⁹⁹.

⁹⁷ VOLTINI, Silvia Pires. **Sentidos da vulnerabilidade no contexto da deficiência a partir da análise de decisões judiciais**. 2023. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2023. Disponível em: <https://repositorio.ufmg.br/handle/1843/55177>. Acesso em: 10 ago. 2024. p. 42-46.

⁹⁸ LUNA, Florencia. Elucidating the concept of vulnerability. **The International Journal of Feminist Approaches to Bioethics**, Toronto, v. 2, n. 1, p. 120-138, 2009. p. 127.

⁹⁹ LUNA, Florencia. Elucidating the concept of vulnerability. **The International Journal of Feminist Approaches to Bioethics**, Toronto, v. 2, n. 1, p. 120-138, 2009. p. 127.

“The metaphor of a layer gives the idea of something softer, something that may be multiple and different, and that may be removed layer by layer. It is not a solid and unique vulnerability that exhausts the category; there

Essa abordagem permite considerar que uma pessoa pode estar em uma posição de vulnerabilidade em alguns aspectos e não em outros, e que essas vulnerabilidades podem interagir e se sobrepor. Por exemplo, no ambiente digital, a vulnerabilidade técnica relacionada ao desconhecimento sobre o funcionamento de tecnologias pode se somar à vulnerabilidade informacional, resultante da assimetria de informações entre os usuários e as plataformas digitais¹⁰⁰.

Além disso, essa situação de fragilidade do usuário da Internet pelo desconhecimento e assimetria de informações pode se sobrepor a outras condições sociais – como o pertencimento a um grupo minoritário, marginalizado ou que demanda proteção adicional, como crianças e adolescentes – e representar a sobreposição de “camas de vulnerabilidade” em relação a alguns usuários de Internet, que demandariam a elaboração de mecanismos de proteção específicos, que também devem se pautar pelo respeito e equilíbrio da autonomia e do livre desenvolvimento do sujeito.

Diante disso, a utilização do conceito de vulnerabilidades em camadas, proposta por Luna (2009), oferece uma lente de análise adequada tanto para a compreensão de fenômenos de risco ou de dano *online* quanto para a proposição de mecanismos de proteção à personalidade na Internet. No caso da manifestação digital do discurso de ódio, por exemplo, identifica-se, em primeiro lugar, que os alvos desse tipo de violência são grupos vulneráveis, cuja identidade ou características são utilizadas para o insulto e a instigação à discriminação. A partir da consideração de que não é possível definir um rol exaustivo de que grupos seriam vulneráveis, e que esta análise depende do contexto social no qual os sujeitos estão inseridos, o conceito de vulnerabilidade em camadas pode ser utilizado para a própria definição do que seria o discurso de ódio. Esta proposta está em consonância com a já identificada importância do contexto para a caracterização do discurso de ódio, tanto para a identificação de suas manifestações concretas nas plataformas digitais, quanto para o surgimento de um ambiente social mais propício à proliferação do discurso de ódio:

[...] estamos em uma era de maior instabilidade social criada por uma maior disparidade de renda [...]. A internet é um campo de batalha natural para quaisquer

might be different vulnerabilities, different layers operating. These layers may overlap: some of them may be related to problems with informed consent, others to social circumstances. The idea of layers of vulnerability gives flexibility to the concept of vulnerability” (LUNA, 2009, p. 127, tradução própria).

¹⁰⁰MARQUES, Claudia Lima; MUCELIN, Guilherme. Vulnerabilidade na era digital: um estudo sobre os fatores de vulnerabilidade da pessoa natural nas plataformas, a partir da dogmática do Direito do Consumidor. **Civilistica.com**, Rio de Janeiro, v. 11, n. 3, p. 1-30, dez. 2022. Disponível em: <https://civilistica.emnuvens.com.br/redc/article/view/872>. Acesso em: 7 nov. 2023. p. 07-14.

pontos de ruptura aos quais chegemos, e também é um mecanismo de vigilância, controle e propaganda. [...] Isso cria um diálogo cada vez mais inflamado, debates polarizados, etc. [...]

Infelizmente, a prática de *trolling* [prática de utilizar a Internet de maneira violenta ou de insultar, com a finalidade de perturbar ou irritar alguém] se tornou uma expressão de uma tendência social mais ampla em várias nações, voltada para o sectarismo beligerante no debate público, com ataques particulares voltados a mulheres e minorias étnicas, religiosas ou sociais¹⁰¹.

Além disso, a infraestrutura das plataformas digitais, que representa uma camada de vulnerabilidade digital aos usuários, pode amplificar o alcance de conteúdos discriminatórios e nocivos, como já identificado em estudos realizados sobre o algoritmo de recomendação do Youtube¹⁰². Não há transparência dos mecanismos de direcionamento de conteúdo personalizado que possa viabilizar a compreensão dos usuários sobre o processo ao qual estão submetidos. Além disso, a vulnerabilidade no ambiente digital é acentuada pela falta de transparência das plataformas em relação às suas políticas de moderação de conteúdo e ao funcionamento de seus algoritmos de recomendação. Essa opacidade cria um ambiente onde o discurso de ódio pode prosperar, explorando as vulnerabilidades dos indivíduos que estão, de fato, mais suscetíveis ao dano em função dessas camadas sobrepostas de fragilidade.

Outro fator importante que se soma a esse é a possibilidade de anonimato e de distanciamento da vítima por parte dos sujeitos que perpetuam o discurso de ódio. O contato indireto entre os usuários e a possibilidade de escapar da responsabilidade pelas ações violentas por meio da utilização de contas anônimas ou falsas propicia um ambiente no qual os usuários se sentem mais confortáveis para perpetuar comportamentos que seriam facilmente condenáveis em interações face a face, o que também representa mais um fator de vulnerabilidade para as vítimas do discurso de ódio. Conforme um relatório de pesquisa divulgado pelo Pew Research Center¹⁰³, o anonimato era visto como um grande potencial da

¹⁰¹ RAINIE, Lee; ANDERSON, Janna; ALBRIGHT, Jonathan. **The future of free speech, trolls, anonymity and fake news online**. Pew Research Center, Washington, 29 mar. 2017. Disponível em: <https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>. Acesso em: 25 mar. 2024. Tradução própria.

“We are in an era of greater social instability created by greater wealth disparity [...]. The internet is the natural battleground for whatever breaking point we reach to play out, and its also a useful surveillance, control, and propaganda tool. [...] That will create even more inflammatory dialogue, flamewars, polarized debates, etc. Unfortunately, I see the present prevalence of trolling as an expression of a broader societal trend across many developed nations, towards belligerent factionalism in public debate, with particular attacks directed at women as well as ethnic, religious, and sexual minorities” (Rainie; Anderson; Albright, 2017, s.p.).

¹⁰² MCCROSKY, Jesse; GEURKINK, Brandi. **Youtube Regrets: A crowdsourced investigation into Youtube’s recommendation algorithm**. Mozilla Foundation, Califórnia, 2021. Disponível em: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Acesso em: 30 out. 2023.

¹⁰³ RAINIE, Lee; ANDERSON, Janna; ALBRIGHT, Jonathan. **The future of free speech, trolls, anonymity and fake news online**. Pew Research Center, Washington, 29 mar. 2017. Disponível em: <https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>. Acesso em: 25 mar. 2024. Tradução própria.

Internet em seus primórdios, pois viabilizaria um ambiente de maior liberdade e criatividade. Porém, no contexto atual, é vista como um fator que propicia comportamentos e discursos inadequados, e um dos maiores desafios para a regulação dos ambientes digitais é o equilíbrio entre a proteção ao anonimato e à privacidade dos usuários e a responsabilização por comportamentos abusivos no meio virtual.

Por fim, a prática de discurso de ódio digital pode representar um fator de aprofundamento de vulnerabilidades, tanto no âmbito digital quanto no *offline*. A incitação à discriminação e à violência, que é conteúdo essencial da prática do discurso de ódio, tem como consequência o aprofundamento de desigualdades sociais, da exclusão e da marginalização de grupos que já se encontravam em situação de vulnerabilidade, seja por meio da prática ou ameaça de violência física, ou pela prática de comportamentos reiterados que provocam a exclusão de espaços virtuais de debate público. Essas formas de dano, tanto *online* quanto *offline*, retroalimentam um ciclo de violência e exclusão que é essencialmente fundado nas vulnerabilidades.

Dessa maneira, a necessidade de proteção da pessoa vulnerável oferece uma lente de análise protetiva e adequada às manifestações complexas do problema da discriminação e do discurso de ódio na Internet, tanto para a compreensão e delimitação do problema quanto para a elaboração de políticas voltadas para o combate ao problema, que estejam voltadas para a maximização da proteção à personalidade em consonância com o respeito ao desenvolvimento da autonomia pessoal.

4 CARACTERIZAÇÃO E REGULAÇÃO DO DISCURSO DE ÓDIO: UM ESTUDO DE CASO DA PLATAFORMA YOUTUBE

4.1 Contextualização: o ecossistema digital do Youtube

Fundado pela Google em 2005, o Youtube é a segunda rede social mais utilizada do mundo, de acordo com a métrica de acessos mensais dos usuários¹⁰⁴. Já no Brasil, segundo pesquisas realizadas no primeiro semestre de 2024, a plataforma é o segundo domínio mais acessado no país, com alcance potencial de 144 milhões de pessoas¹⁰⁵. Tanto a grande

¹⁰⁴ MOST popular social networks worldwide as of April 2024, ranked by number of monthly active users (in millions). Statista, ago. 2024. Disponível em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 12 ago. 2024.

¹⁰⁵ KEMP, Simon. **Digital 2024: Brazil**. Data Reportal, 23 fev. 2024. Disponível em: <https://datareportal.com/reports/digital-2024-brazil>. Acesso em: 12 jun. 2024.

popularidade da plataforma quanto o fato de que pesquisas recentes a considerem como uma rede social e não um serviço de vídeos sob demanda podem ser surpreendentes, especialmente dada a prevalência de estudos de caso sobre outras plataformas em pesquisas sobre redes sociais¹⁰⁶.

A caracterização do Youtube como uma rede social pode provocar questionamentos, já que a plataforma pode ser inicialmente vista como um mero serviço de disponibilização de vídeos. Apesar da variação de definições formais, considera-se que as redes sociais possuem algumas características e funcionalidades comuns, dentre as quais: serem aplicações baseadas na Internet, nas quais há tanto conteúdo quanto perfis gerados pelos usuários, e serem formadas por conexões entre os perfis¹⁰⁷.

Diante disso, percebe-se que o Youtube se distingue de serviços de vídeo sob demanda por permitir a interação constante e bidirecional entre criadores de conteúdo e sua audiência, o que caracteriza essencialmente uma rede social. Os usuários não apenas assistem aos vídeos, mas também comentam, compartilham, criam conteúdo de resposta e se engajam em discussões nas seções de comentários.

Munger e Phillips (2022)¹⁰⁸ identificaram que o Youtube possui uma série de particularidades em sua estruturação e funcionamento que o caracterizam como um “mercado da informação”, no qual os criadores de conteúdo se utilizam das ferramentas próprias disponibilizadas pela plataforma, como o algoritmo de recomendação e a monetização de vídeos, para disputar pela atenção de uma audiência com tempo limitado. Os pesquisadores destacam que a forma de comunicação por vídeo é uma forma de consumo de conteúdo mais natural para o ser humano do que, por exemplo, a comunicação por escrito, o que é um fator adicional que justifica a grande popularidade da plataforma entre o público mundial. Existiria, também, uma demanda de consumo da informação pelo público, que é definida por elementos como hábitos de consumo do conteúdo e circunstâncias sociais e econômicas.

Para os autores (2022), esses elementos definem o Youtube como um importante fórum de debates políticos e sociais e de formação de opinião:

¹⁰⁶ TUFEKCI, Zeynep. Big questions for social media big data: representativeness, validity and other methodological pitfalls. **ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media**, 2014. Disponível em: <https://arxiv.org/pdf/1403.7400>. Acesso em: 12 ago. 2024.

¹⁰⁷ OBAR, Johnatan A.; WILDMAN, Steve. Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. **Telecommunications Policy**, Oxford, v. 39, n. 9, p. 1-21, 2015. Disponível em: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2927584_code962598.pdf?abstractid=2647377&mirid=1&type=2. Acesso em: 15 ago. 2024. p. 07-10.

¹⁰⁸ MUNGER, Kevin; PHILLIPS, Joseph. Right-Wing YouTube: a supply and demand perspective. **The International Journal of Press/Politics**, Califórnia, v. 27, n. 1, p. 186-219, 2022. Disponível em: <https://journals.sagepub.com/doi/10.1177/1940161220964767>. Acesso em: 21 ago. 2024. p. 193.

O advento da Internet colocou as massas dentro deste ciclo de inovação. Em primeiro lugar, as redes de discussões políticas extrapolam a comunidade local ou a família e agora ocorrem *online* e envolvem pessoas do mundo inteiro. Em segundo lugar, a limitação à capacidade de produção de mídia foi radicalmente reduzida, e as pessoas cada vez mais consomem informação de conteúdo gerado pelos usuários. O Youtube se encontra na intersecção desses desenvolvimentos. Pesquisas na área da comunicação reconhecem o Youtube como um importante espaço de formação do discurso, um local para grupos minoritários, e de troca entre o criador de conteúdo e sua audiência¹⁰⁹.

Logo, o alcance do Youtube se estende para além do entretenimento, e a plataforma assume um importante papel para a disseminação de informações, debates políticos e a expressão de opiniões diversas. No entanto, essa influência também levanta preocupações sobre a disseminação de conteúdo nocivo, como o discurso de ódio e conteúdo violento e extremista.

Um estudo¹¹⁰ conduzido sobre o comportamento de usuários do Youtube que se engajam com conteúdo extremista identificou a existência de uma “progressão de radicalização” (*radicalization pipeline*). A plataforma não somente hospeda conteúdo de natureza extremista e controversa, o que é amplamente constatado, mas, por meio da análise da atividade de usuários nos comentários de vídeos e canais, identificou-se que a forma de consumo desse tipo de conteúdo facilita a progressão de usuários de conteúdos mais moderados para mais extremos. Isso ocorre em razão tanto em razão das conexões entre os usuários quanto em razão do sistema de recomendações do Youtube, que, por meio do algoritmo de sugestões, inadvertidamente promoveria a exposição dos usuários a ideias extremistas ao sugerir conteúdos cada vez mais polarizados e radicais.

Nesse contexto, a moderação de conteúdo se destaca como uma ferramenta essencial para mitigar os riscos associados à propagação de material nocivo, como a difusão de discurso de ódio. Contudo, a implementação eficaz de políticas de moderação de conteúdo enfrenta

¹⁰⁹ MUNGER, Kevin; PHILLIPS, Joseph. Right-Wing YouTube: a supply and demand perspective. **The International Journal of Press/Politics**, Califórnia, v. 27, n. 1, p. 186-219, 2022. Disponível em: <https://journals.sagepub.com/doi/10.1177/1940161220964767>. Acesso em: 21 ago. 2024. p. 190.

“The advent of the World Wide Web has brought the masses into this cycle of innovation. First, political discussion networks, long the purview of one’s neighborhood and extended family, now occur online with people throughout the world. Second, the gatekeeping process for being able to produce media has been radically reduced, and people increasingly get their news from user-generated content. YouTube lies at the intersection of these developments. Communication research, largely qualitative, has long recognized YouTube as an important site of discourse, a site for counter-publics, and of interplay/interchangeability between creator and audience” (Munger; Phillips, 2022, p. 190).

¹¹⁰ RIBEIRO, Manoel Horta; OTTONI, Raphael et al. Auditing Radicalization Pathways on YouTube. In: PROCEEDINGS of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: Association for Computing Machinery, 2020. (FAT* ’20), p. 131–141. ISBN 9781450369367. DOI: 10.1145/3351095.3372879.

desafios significativos. A adoção de políticas claras e rigorosas contra o discurso de ódio deve ser acompanhada por mecanismos que garantam transparência no processo de moderação e na operação dos algoritmos de recomendação. Por exemplo, a utilização de etapas automatizadas de identificação de conteúdo nocivo deve ser feita de forma que minimize o risco de remoção de conteúdos legítimos e de manutenção de conteúdos violadores no ar.

Além disso, Munger e Phillips¹¹¹ argumentam que é imperativo que essas práticas de moderação considerem a natureza complexa do conteúdo e do ecossistema no qual os usuários interagem. A moderação não deve ser vista apenas como um meio de remover conteúdos prejudiciais, mas também como uma parte integrante de um esforço mais amplo para garantir que o Youtube permaneça um espaço de diálogo produtivo, diante da ampla potencialidade da plataforma como um espaço de formação da opinião pública. Diante das constatações de que a radicalização dos usuários não é apenas uma possibilidade teórica, mas um fenômeno mensurável que ocorre devido, em parte, à estrutura algorítmica da plataforma, é necessária uma reavaliação crítica de como o algoritmo de recomendação influencia a exposição dos usuários a tais conteúdos.

Em relação a isso, uma pesquisa conduzida pela Fundação Mozilla¹¹² que coletou informações sobre a experiência de usuários que haviam entrado em contato com conteúdo nocivo ou danoso, concluiu que essa exposição decorria significativamente das recomendações automatizadas da plataforma. O conteúdo assim classificado pelos voluntários da pesquisa apresentava uma boa performance em termos de popularidade, e o contato da maioria dos voluntários (71%) com esses vídeos ocorreu de forma involuntária, por meio das recomendações automatizadas¹¹³. Essas conclusões evidenciam a importância do ranqueamento automatizado como forma de moderação de conteúdo na plataforma e de definição da experiência do usuário. Diante desses resultados, a fundação recomenda que o Youtube e as demais plataformas adotem medidas que aumentem a transparência dos sistemas de recomendação, inclusive com a viabilização de auditorias independentes, e que favoreçam a autonomia do usuário, possibilitando que ele compreenda o funcionamento do sistema de recomendação, a utilização de seus dados para o ranqueamento personalizado de conteúdo, e

¹¹¹ MUNGER, Kevin; PHILLIPS, Joseph. Right-Wing YouTube: a supply and demand perspective. **The International Journal of Press/Politics**, Califórnia, v. 27, n. 1, p. 186-219, 2022. Disponível em: <https://journals.sagepub.com/doi/10.1177/1940161220964767>. Acesso em: 21 ago. 2024. p. 198.

¹¹² MCCROSKY, Jesse; GEURKINK, Brandi. **Youtube Regrets**: A crowdsourced investigation into Youtube's recommendation algorithm. Mozilla Foundation, Califórnia, 2021. Disponível em: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Acesso em: 30 out. 2023.

¹¹³ MCCROSKY, Jesse; GEURKINK, Brandi. **Youtube Regrets**: A crowdsourced investigation into Youtube's recommendation algorithm. Mozilla Foundation, Califórnia, 2021. Disponível em: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Acesso em: 30 out. 2023. p. 13.

inclusive a desabilitação das recomendações personalizadas, caso assim deseje¹¹⁴. Porém, a adoção dessas medidas pelas plataformas e a criação de políticas públicas por autoridades estatais voltadas à natureza desses riscos é incipiente, quando não inexistente, e contribui para a manutenção de um ambiente de criação e disseminação de conteúdos nocivos, com riscos concretos *online* e *offline* sobretudo em relação a populações vulneráveis.

4.2 Política da plataforma para o discurso de ódio

O Youtube adota diretrizes para coibir a difusão de discurso de ódio, publicadas nas “regras da comunidade”¹¹⁵. As regras foram atualizadas em 2019 para aprimorar o combate ao discurso de ódio, “honrar a responsabilidade [da plataforma] e proteger a comunidade de conteúdo nocivo”, e “limitar a disseminação de conteúdo extremista e violento *online* em parceria com legisladores e a sociedade civil de todo o mundo”, conforme comunicado oficial do próprio Youtube¹¹⁶. A política oficial se basearia em quatro pilares: (i.) remoção de conteúdo violador; (ii.) promoção de conteúdo autoritativo; (iii.) redução do alcance de conteúdo *borderline* (conteúdo que não é explicitamente ilegal, mas ainda assim promove narrativas extremistas, incita a violência ou têm um efeito polarizador¹¹⁷); e (iv.) beneficiar criadores de conteúdo confiáveis.

Conforme a publicação oficial da plataforma, não é admitido conteúdo que promova “violência ou ódio contra indivíduos” de grupos protegidos, cujos atributos são definidos previamente. São estes: “idade, classe social, deficiência, etnia, identidade ou expressão de gênero, nacionalidade, raça, *status* de imigrante, religião, sexo ou gênero, orientação sexual, vítimas de um evento violento e seus descendentes, e veteranos [de guerra]”¹¹⁸.

¹¹⁴ MCCROSKY, Jesse; GEURKINK, Brandi. **Youtube Regrets**: A crowdsourced investigation into Youtube’s recommendation algorithm. Mozilla Foundation, Califórnia, 2021. Disponível em: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Acesso em: 30 out. 2023. p. 25-27.

¹¹⁵ YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

¹¹⁶ YOUTUBE TEAM. **Our ongoing work to tackle hate**. San Bruno: Youtube Official Blog, maio 2019. Disponível em: <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/>. Acesso em: 6 ago. 2024.

¹¹⁷ COMISSÃO EUROPEIA. **Call for participants**: RAN C&N meeting on How to deal with borderline content (related to hate speech, meme culture, humour etc.) from the perspective of public trust? Face-to-face meeting, Rotterdam, The Netherlands on 16 and 17 April 2024. European Commission: Migration e Home Affairs, News Announcement, 01 mar. 2024. Disponível em: https://home-affairs.ec.europa.eu/news/call-participants-ran-cn-meeting-how-deal-borderline-content-related-hate-speech-meme-culture-humour-2024-03-01_en. Acesso em: 6 ago. 2024.

¹¹⁸ YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

- O Youtube também determina previamente que tipo de conteúdo se caracteriza como discurso de ódio para a finalidade de adequação à política da comunidade. Este é definido como a utilização de qualquer produto ou funcionalidade do Youtube (vídeos, descrições, comentários, transmissões ao vivo, links para sites externos, e outros) com a finalidade de¹¹⁹:
- Encorajar violência contra os grupos protegidos, incluindo ameaças implícitas;
- Incitar ódio contra grupos protegidos;
- Desumanizar indivíduos ou grupos;
- Exaltar violência contra indivíduos ou grupos baseando-se em seu *status* de protegido;
- Utilizar ofensas ou estereótipos raciais ou religiosos que incitam ou promovem ódio baseando-se nas características do grupo protegido;
- Alegar que indivíduos ou grupos são fisicamente ou mentalmente inferiores, deficientes ou doentes por serem membros do grupo protegido;
- Promover a supremacia odiosa ao alegar a superioridade de um grupo sobre outro com *status* protegido para justificar violência, discriminação, segregação ou exclusão;
- Fazer alegações conspiratórias de que indivíduos ou grupos são maus, corruptos ou maliciosos por pertencerem a uma categoria protegida;
- Negar ou minimizar eventos violentos importantes e bem-documentados; e
- Atacar indivíduos ou grupos em razão da atração emocional, romântica ou sexual a outras pessoas.

Como exceção, são mencionados vídeos que incluem o discurso de ódio com caráter educativo, científico ou artístico, ou que tem o objetivo de condenar, refutar, se opor ou satirizar esta prática. Como exemplo, menciona-se documentários sobre grupos de ódio em tom educativo, sem a promoção de ideais de violência, ou materiais sobre eventos históricos, como a Segunda Guerra Mundial ou a evolução das teorias sobre a espécie humana¹²⁰.

Em relação à identificação do conteúdo nocivo, o Youtube adota um mecanismo que permite que os próprios usuários denunciem vídeos, canais ou demais conteúdos que são inapropriados por violarem as regras da comunidade, inclusive aqueles que promovem discurso de ódio¹²¹. Ademais, o conteúdo pode ser vistoriado por especialistas, como indivíduos, ONGS, entidades governamentais ou acadêmicas. A plataforma também informa

¹¹⁹ YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

¹²⁰ YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

¹²¹ YOUTUBE. **Report inappropriate videos, channels, and other content on YouTube**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2802027>. Acesso em: 6 ago. 2024.

contar com algoritmos de *machine learning* voltados para a detecção automática de conteúdo que potencialmente viola as regras da comunidade. Algumas das decisões são inteiramente automáticas, mas outras são encaminhadas para moderadores humanos que conferem os metadados e o conteúdo, e podem optar por removê-lo, restringir seu acesso (por exemplo, limitando a idade), ou mantê-lo no ar¹²².

Por fim, também há punições previstas para os casos de violação às regras de combate ao discurso de ódio. Segundo a própria plataforma, em “alguns casos raros” o Youtube pode remover conteúdo ou aplicar penalidades quando um criador de conteúdo apresentar comportamento nocivo. Novamente, são listadas as atitudes que dão ensejo à aplicação das punições¹²³:

- Encorajar repetidamente comportamento abusivo dos espectadores;
- Repetidamente insultar, atacar ou abusar, em múltiplos *uploads*, um grupo com base em suas características protegidas;
- Expor um grupo protegido a riscos de dano físico em razão do contexto político ou social local; e
- Criar conteúdo que prejudica o ecossistema do Youtube por, de forma persistente, incitar hostilidade contra um grupo protegido com o objetivo de obter vantagem financeira pessoal.

Caso se identifique que o conteúdo viola a política do Youtube, ele é derrubado e o criador é informado via *e-mail*. Além disso, há um escalonamento de advertências para o infrator: se for a primeira violação, o criador deve participar em um treinamento sobre as políticas da comunidade. Em seguida, após o decurso de 90 dias sem infrações, a advertência expira e é removida. Se é cometida uma nova infração, a advertência é mantida, e o usuário recebe uma marcação, ou *strike*. Se 3 *strikes* forem acumulados em uma janela de 90 dias, o canal é derrubado¹²⁴.

O Youtube divulga, em relatórios de transparência trimestrais, as estatísticas de remoção de conteúdo por violação às políticas da comunidade¹²⁵. No total, compreendendo

¹²²YOUTUBE VIWERS. **The life of a flag**. San Bruno: Youtube, 2018. Disponível em: <https://www.youtube.com/watch?v=WK8qRNSmhEU>. Acesso em: 6 ago. 2024.

¹²³YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

¹²⁴YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

¹²⁵GOOGLE. **Relatório de Transparência**. 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: <https://transparencyreport.google.com/youtubepolicy/removals>. Acesso em: 7 ago. 2024.

GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: <https://transparencyreport.google.com/youtube->

todas as remoções de conteúdo por violação a qualquer uma das políticas da comunidade, entre janeiro de 2024 e março de 2024¹²⁶, foram removidos 15.799.880 canais, contendo 103.815.590 vídeos. Além disso, foram removidos 8.295.304 vídeos individuais, sem incluir aqueles derrubados pela exclusão do canal, e 1.443.821.162 comentários.

Estima-se que haja uma média de 114 milhões de canais ativos no Youtube¹²⁷, portanto, a porcentagem de canais removidos por violação às políticas da comunidade é aproximadamente 13% no período estudado. Em relação aos vídeos, a estimativa é de que haja aproximadamente 14 bilhões de *uploads*¹²⁸. Assim, a quantidade total de vídeos derrubada em razão da moderação de conteúdo (112.110.894 – soma de vídeos individuais e vídeos de canais excluídos) é de 0,8% de todos os vídeos existentes no Youtube. Não há a divulgação pela plataforma da quantidade oficial de canais, vídeos ou comentários, e as estimativas de conteúdo total utilizadas foram feitas por terceiros.

Em relação à identificação do conteúdo nocivo, ou *flagging*, percebe-se que a vasta maioria dos vídeos removidos foi identificada pelo mecanismo automático de moderação do Youtube. Somente 6 vídeos em um universo de mais de 8 milhões foram removidos por detecção inicial de autoridades estatais. Os dados disponibilizados dizem respeito aos vídeos denunciados e efetivamente derrubados, e não é possível saber qual é o número total de denúncias feitas por autoridades estatais que não resultou na remoção do conteúdo, ou se os 6 vídeos removidos dizem respeito a todas as ordens de derrubada por parte de autoridades estatais recebidas no trimestre analisado.

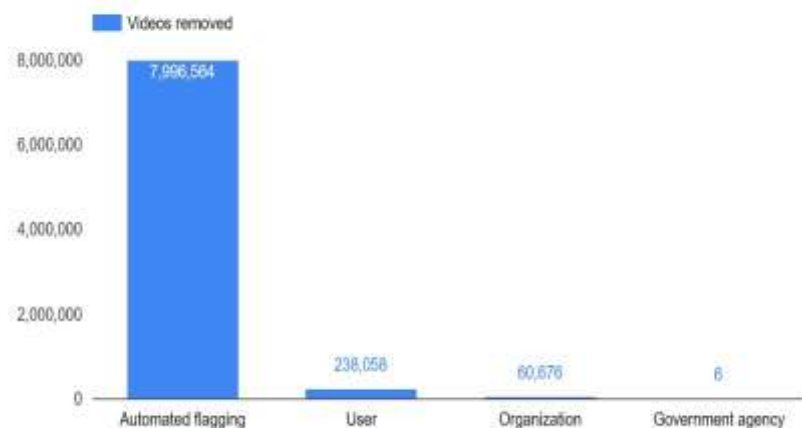
policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=videos_by_country&content_by_flag=period:2024Q1;exclude_automated=all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:. Acesso em: 7 ago. 2024.

¹²⁶ Período mais recente para consulta disponível em agosto de 2024.

¹²⁷ WYZOL. **Youtube Stats**: 2025. Disponível em: <https://www.wyzowl.com/youtube-stats/>. Acesso em: 26 maio 2025.

¹²⁸ MCGRADY, Ryan. **What we discovered on Deep Youtube**. The Atlantic, Washington, 26 jan. 2024. Disponível em: <https://www.theatlantic.com/technology/archive/2024/01/how-many-videos-youtube-research/677250/>. Acesso em: 26 maio 2025.

Figura 2 – Origem da identificação do conteúdo derrubado

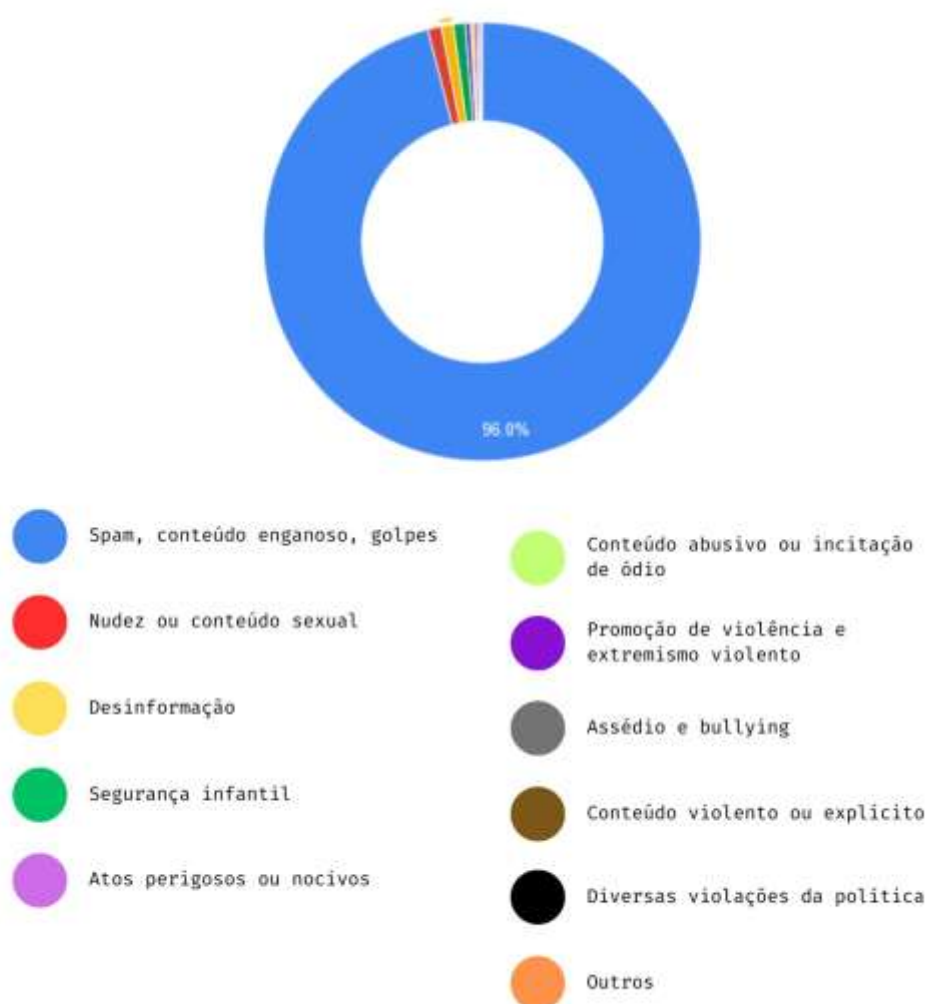


Fonte: Google, 2024¹²⁹

Outro ponto relevante do relatório de transparência é a motivação para a derrubada de conteúdo. Conforme a própria classificação feita pelo Youtube, a retirada do ar de canais ou de vídeos por “conteúdo abusivo ou incitação de ódio” e “promoção de violência e extremismo violento”, que caracterizam o discurso de ódio, segundo a própria classificação feita pelo Youtube, representam porcentagens pequenas. Somente 0,4% dos canais foram derrubados pela violação às políticas de combate ao ódio e à promoção de violência (0,2% para cada categoria), o que representa 74.279 canais. Já em relação aos vídeos, a incitação de ódio e a promoção de violência somam 5.6% do total de derrubadas (3,7% e 1.9%, respectivamente), totalizando a soma de 468.867.

¹²⁹GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

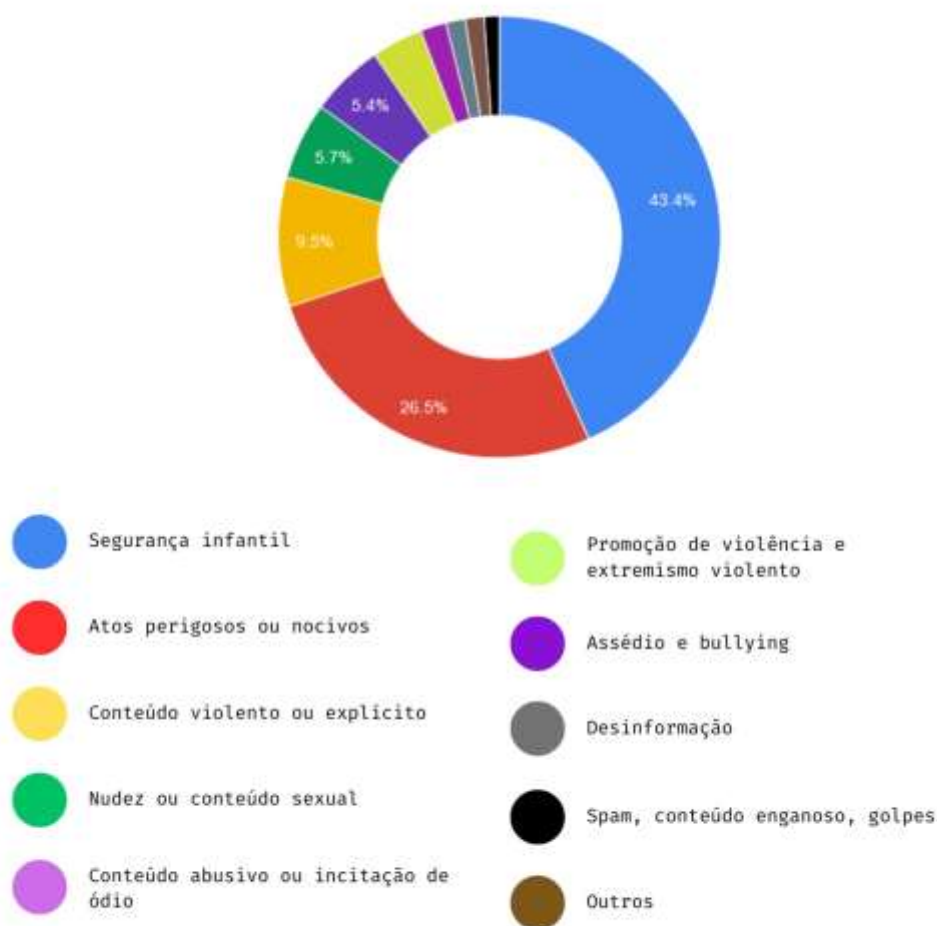
Figura 3 – Derrubada de canais



Fonte: Google, 2024¹³⁰

¹³⁰ GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

Figura 4 – Derrubada de vídeos



Fonte: Google, 2024¹³¹

A quantidade total de vídeos e canais derrubados não é insignificante: no lapso de três meses, foram removidos mais de 74 mil canais e quase 470 mil vídeos que promoviam o discurso de ódio. Por outro lado, e considerando as estimativas da quantidade total de vídeos

¹³¹ GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

(14 bilhões) e canais (cerca de 114 milhões) ativos na plataforma, pode-se deduzir que a quantidade total de conteúdo que promove o discurso de ódio no Youtube seja significativamente superior à quantidade de vídeos e canais derrubados, notadamente em razão das políticas que a empresa adota para o combate a este problema. A título de comparação, a porcentagem de canais derrubados no período pelas causas identificadas como correlatas ao discurso de ódio é de aproximadamente 0.06% do total, enquanto a porcentagem de vídeos é de 0.003%, ambas proporções ínfimas diante do universo de conteúdo existente na plataforma.

Por fim, o relatório de transparência¹³² traz dados sobre a localização do IP onde foi feito o *upload* dos vídeos removidos. Em relação a todas as remoções agregadas, o Brasil ocupa o segundo lugar, com 776.923 vídeos retirados do ar entre janeiro e março de 2024, atrás da Índia, com um total de 2.618.760 vídeos. O relatório ressalta que pode haver distorções nos dados pela utilização de VPNs ou *proxy servers*¹³³. Ainda assim, os dados indicados pelo relatório e a popularidade do Youtube entre o público brasileiro indica a relevância da questão no âmbito nacional.

4.3 Desafios e perspectivas para o combate ao discurso de ódio no Youtube

Inicialmente, observa-se que a definição do que é o discurso de ódio pela política do Youtube é consideravelmente mais restritiva do que definições mais amplas, como a adotada pela ONU¹³⁴ e pela Comissão Europeia¹³⁵, que não limitam previamente que atos devem ser considerados como violentos ou que grupos estão sujeitos ao discurso de ódio. A plataforma proíbe conteúdo que promova violência ou ódio contra indivíduos com base em atributos pré-definidos, mas essa definição pode deixar de fora formas mais sutis de discurso de ódio que

¹³² GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

¹³³ Com a utilização de VPNs ou de *proxy servers*, o usuário pode alterar a localização do seu acesso à Internet. Assim, um usuário localizado no Brasil, mas que utilize uma VPN com endereço na Índia e que tenha seu vídeo derrubado, será computado dentro da estatística da Índia.

¹³⁴ ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **UN strategy and plan of action on hate speech**. ONU, Nova Iorque, maio 2019. Disponível em: https://www.un.org/sites/un2.un.org/files/action_plan_on_hate_speech_en.pdf. Acesso em: 12 ago. 2024.

¹³⁵ JOROUVÁ, Vera. **Code of Conduct for Illegal online hate speech**: questions and answers. European Commission, Bruxelas, jun. 2016. Disponível em: https://ec.europa.eu/newsroom/document.cfm?doc_id=41844. Acesso em: 6 ago. 2024.

não se enquadram nas categorias listadas. A definição prévia de atributos protegidos de grupos sociais e de conteúdos proibidos parece ter como objetivo assegurar aos usuários do Youtube que a plataforma preza pela livre expressão e que a remoção de conteúdo considerado nocivo é uma exceção. Contudo, essa escolha por parte da plataforma representa um obstáculo adicional ao combate ao discurso de ódio, tendo em vista que contraria as considerações teóricas trazidas pela teoria das vulnerabilidades sobre a proteção do sujeito na Internet.

Considerando que o discurso de ódio virtual está intrinsecamente ligado às vulnerabilidades digitais, a abordagem do Youtube na proteção desses grupos se aproxima do posicionamento teórico criticado por Luna (2009)¹³⁶, ao criar rótulos imutáveis de vulnerabilidades, que não dão margem para a identificação de situações de fragilidade que venham a surgir pela alteração do contexto social, ou para o reconhecimento de graus de vulnerabilidade distintos dentro dos próprios grupos protegidos pela sobreposição de camadas. O emprego de cláusulas abertas e definições demasiado fluidas do que é o discurso de ódio é criticado por alguns teóricos¹³⁷, que tem razão ao afirmar que essas tentativas de definição vagas podem dificultar a operacionalização das políticas de combate ao discurso de ódio, que se baseiam em uma primeira abordagem de detecção automatizada nas grandes plataformas. Porém, o estudo de caso sobre as políticas do Youtube também indica um alerta na direção contrária: a adoção de critérios demasiadamente restritivos e exaustivos para a caracterização do discurso de ódio pode causar lacunas na proteção a grupos vulneráveis e se tornar obsoleta diante da constante alteração do contexto social e das formas de comunicação.

Um outro aspecto relacionado aos desafios para o combate efetivo ao discurso de ódio no Youtube é a afirmação da plataforma de que a derrubada de conteúdo que viola a política de discurso de ódio ocorreria somente em “alguns casos raros”. Porém, não há informação adicional acerca de que elementos caracterizariam esses “casos raros”, o que leva à indicação de que a decisão sobre o que configura ou não um caso grave de discurso de ódio é um ato discricionário da plataforma. A falta de transparência sobre esses critérios impede que haja o controle pelos usuários ou pela sociedade em geral sobre a efetiva implementação da filosofia que a plataforma alega ter em relação à manutenção de um ambiente digital saudável e igualitário. Afinal, é no mínimo contraditório que o Youtube reafirme o seu

¹³⁶ LUNA, Florencia. Elucidating the concept of vulnerability. **The International Journal of Feminist Approaches to Bioethics**, Toronto, v. 2, n. 1, p. 120-138, 2009.

¹³⁷ HIETANEN, Mika; EDDEBO, Johan. Towards a definition of hate speech with a focus on online contexts. Disponível em: **Journal of Communication Inquiry**, Londres, v. 47, n. 4, p. 441-458, out. 2022. <https://journals.sagepub.com/doi/epub/10.1177/01968599221124309>. Acesso em: 7 ago. 2024.

compromisso com o combate ao discurso de ódio, mas depois exponha que apenas alguns casos raros de manifestações desse tipo de violência serão retirados do ar. Este posicionamento, aliado à restrição do que é considerado discurso de ódio na plataforma, cria uma situação propícia à manutenção de conteúdos nocivos no ar.

Um indício desse cenário são os dados disponibilizados nos relatórios de transparência, que revelam a baixa porcentagem de derrubadas de canais e vídeos pela motivação de “conteúdo abusivo ou incitação de ódio” ou “promoção de violência e extremismo violento” (Figuras 3 e 4)¹³⁸. Essa situação pode ser atribuída à existência de um alto número de canais e contas que se caracterizam como *spam* e podem ser facilmente identificados e removidos pelo processo de moderação automática de conteúdo, e da política ativa que o Youtube adota para o combate a práticas enganosas e que prejudiquem a comunidade¹³⁹. Porém, a grande disparidade entre a motivação para a derrubada de conteúdo e a ampla constatação por pesquisadores da forte presença do discurso de ódio no Youtube também são um indicador de que as políticas mais restritivas da plataforma em relação a essa questão permitem que essas manifestações de violência se propaguem de forma desregulada no ambiente digital.

No Brasil, conforme analisado, não há uma legislação específica sobre a caracterização do discurso de ódio *online*, ou que defina uma gramática comum de responsabilização das plataformas que leve à implementação de políticas de uso voltadas ao combate uniformizado do problema. O Marco Civil da Internet (Lei nº 12.965/2014) prevê, em seu artigo 19, que o provedor de aplicações da internet será responsabilizado por danos decorrentes de conteúdo postado por usuários caso desobedeça a ordem judicial específica de remoção, excetuando a hipótese de violação da intimidade decorrente de divulgação de conteúdo sexual (artigo 21)¹⁴⁰. O Projeto de Lei nº 2.630/2020 apresentava como uma de suas principais propostas o estabelecimento de uma regulação uniforme e mais abrangente da responsabilidade das plataformas pelo conteúdo nelas veiculado. Embora não houvesse uma

¹³⁸ GOOGLE. **Relatório de Transparência**. Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

¹³⁹ YOUTUBE. **Políticas contra spam, práticas enganosas e golpes**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2801973?hl=pt-BR>. Acesso em: 6 ago. 2024.

¹⁴⁰ BRASIL. 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 abr. 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 12 ago. 2024.

disposição específica no texto original sobre o discurso de ódio, um dos objetivos da proposta é de promover “a busca por maior transparência das práticas de moderação de conteúdos postados por terceiros em redes sociais, com a garantia do contraditório e da ampla defesa” (artigo 4º, inciso III)¹⁴¹. Ademais, o PL buscava estabelecer diretrizes gerais para o procedimento de moderação de conteúdo (Seção IV) e para a criação de instituições de “autorregulação regulada” (artigo 30) dentro das plataformas, em uma abordagem que se aproxima das propostas de correção. A tramitação do PL nº 2.630/2020 foi marcada por debates e controvérsias, e a aprovação do texto original no atual cenário político é considerada como improvável, após 4 anos da apresentação da proposta¹⁴². Assim, o cenário regulatório da questão no Brasil ainda se aproxima da autorregulação das plataformas.

Apesar das particularidades do problema da regulação do ambiente digital e da impossibilidade da aplicação de uma lógica tradicional que busque abranger todas as soluções para um problema¹⁴³, a análise das formas de combate ao discurso de ódio adotadas pelo Youtube aponta para a necessidade da instituição, pelo poder público, de diretrizes comuns que possam ser operacionalizadas pelas plataformas digitais para o combate de manifestações de violência *online*. A discricionariedade das empresas para definir o que é o discurso de ódio e os casos em que seria devida a remoção de conteúdo tem como consequência o estabelecimento de um ambiente digital desprotegido, propício para o aprofundamento de vulnerabilidades de grupos que já enfrentam desigualdades e marginalizações devido ao contexto social.

Em conclusão, a análise das políticas do YouTube em relação ao discurso de ódio revela uma abordagem que, embora busque equilibrar a liberdade de expressão com a necessidade de moderação, acaba por criar lacunas significativas na proteção dos grupos vulneráveis. A definição restritiva do que constitui discurso de ódio, aliada à falta de transparência sobre os critérios de remoção de conteúdo, resulta em um ambiente digital onde manifestações nocivas podem prosperar de forma desregulada. Esse cenário é exacerbado pela

¹⁴¹VIEIRA, Alessandro. **Projeto de Lei nº 2630/2020**. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. Brasília: Senado Federal, 13 maio 2020. Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&ts=1712173889308&disposition=inline>. Acesso em: 14 ago. 2024.

VIEIRA, Alessandro. **Emenda ao Projeto de Lei nº 2630/2020 - SF/20458.58654-0**. Brasília: Senado Federal, 2020. Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&disposition=inline>. Acesso em: 14 ago. 2024.

¹⁴² RECHMANN, Aline. **Aprovação do PL das Fake News é improvável, diz pesquisa**. Gazeta do Povo, Curitiba, 30 maio 2024. Disponível em: <https://www.gazetadopovo.com.br/república/45-dos-deputados-consideram-improvavel-a-aprovacao-do-pl-das-fake-news/>. Acesso em: 14 ago. 2024.

¹⁴³ KLEINWÄCHTER, Wolfgang; ALMEIDA, Virgílio A.F. The internet governance ecosystem and the rainforest. **IEEE Internet Computing**, Washington, v. 19, n. 2, p. 64–67, 2015. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2015.49>. Acesso em: 12 jun. 2024.

ausência de uma regulação mais rigorosa e uniforme no Brasil, que poderia fornecer diretrizes claras para a atuação das plataformas. A persistência desse modelo de autorregulação, criticado por sua insuficiência em lidar com a complexidade e a mutabilidade das vulnerabilidades digitais, sugere a necessidade urgente de uma intervenção regulatória mais robusta, que leve em consideração a estruturação e o funcionamento próprio do ambiente digital, mas também priorize a proteção a grupos vulneráveis, que é elemento central do desenvolvimento recente do direito privado brasileiro.

5 CONSIDERAÇÕES FINAIS

A regulação de ambientes digitais é um problema intrincado, marcado por dilemas como a complexidade técnica dos objetos, a estruturação do ecossistema digital, a natureza sociotécnica da Internet e as constantes mudanças no ambiente digital, que requerem a adaptabilidade das políticas e dos modelos regulatórios.

Uma das questões que emergem quando se debate a regulação de ambientes digitais é o controle e filtragem de conteúdo produzido por usuários nas plataformas digitais, especialmente de formas de comunicação nocivas e violentas. Para tanto, tais plataformas se valem de mecanismos de moderação de conteúdo para a definição da experiência *online* dos usuários e a implementação de políticas tanto de origem interna quanto de origem externa à empresa. A moderação de conteúdo é uma atividade essencial à própria caracterização da plataforma, e está intrinsecamente ligada ao serviço oferecido por essas empresas.

Porém, a partir de considerações sobre as novas formas de manifestação do discurso de ódio no meio digital, identificam-se potenciais desafios à concretização da proteção da personalidade contra formas de violência, que pode ter reflexos danosos sobretudo a grupos vulneráveis tanto na esfera *online* quanto na esfera *offline*. A dificuldade de definição do que constitui o discurso de ódio e o fato de que suas manifestações no ambiente digital se baseiam fortemente no contexto da produção da comunicação representam empecilhos à implementação de políticas de moderação de conteúdo efetivamente protetivas, à uniformização da lógica de proteção no ambiente digital, e à análise de uma grande quantidade de conteúdo para a detecção de manifestações nocivas. Aportes teóricos como a teoria das vulnerabilidades descrita por Luna (2009), que coloca ênfase no contexto social no qual o indivíduo está inserido, representam possibilidades de análise do problema de forma a maximizar a proteção à personalidade na Internet, com ênfase na dinâmica do caso concreto e sem descartar de início situações que podem vir a configurar vulnerabilidades.

Apesar do desenvolvimento de estudos e pesquisas que indicam para a importância do contexto e a complexidade do fenômeno do discurso de ódio no ambiente digital, ao se analisar as políticas da plataforma Youtube para a contenção desse problema, identifica-se um cenário de proteção restritiva, que é incapaz de conter a disseminação do discurso de ódio na plataforma. A determinação prévia de situações que configurariam o discurso de ódio, que é contrária ao conceito das vulnerabilidades em camadas, a falta de transparência da plataforma quanto à política de determinação de remoção de conteúdo e a adoção de filosofias empresariais que privilegiam a liberdade de expressão em prol da proteção de grupos vulneráveis geram um ambiente propício à disseminação de conteúdo relacionado ao discurso de ódio. Diante da grande popularidade da plataforma e de sua influência para a formação de opinião pública, este cenário de risco aponta para a necessidade de se repensar o modelo regulatório adotado para a questão no Brasil, especialmente diante da adoção pelo direito brasileiro de postulados que prezam pela proteção existencial a grupos vulneráveis.

REFERÊNCIAS

ABBATE, Janet. What and where is the Internet? (Re)defining Internet histories. **Internet Histories: Digital Technology, Culture and Society**, London, v. 1, n. 1-2, p. 8-14, 2017. Disponível em: <https://doi.org/10.1080/24701475.2017.1305836>. Acesso em: 15 jul. 2026.

ALMEIDA, Virgílio; FILGUEIRAS, Fernando; GAETANI, Francisco. Digital governance and the tragedy of the commons. **IEEE Internet Computing**, Washington, v. 24, n. 4, p. 41-46, 2020. Disponível em: <https://ieeexplore.ieee.org/document/9195909>. Acesso em: 15 jul. 2026.

ALMEIDA, Virgílio; FILGUEIRAS, Fernando; DONEDA, Danilo. The ecosystem of digital content governance. **IEEE Internet Computing**, Washington, v. 25, n. 3, p. 13-17, 2021. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2021.3057756>. Acesso em: 15 jul. 2026.

ALTO COMISSÁRIO DAS NAÇÕES UNIDAS PARA OS DIREITOS HUMANOS. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/74/486)**. Nova Iorque: Assembleia Geral das Nações Unidas, out. 2019. Disponível em: <https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech>. Acesso em: 8 ago. 2024.

ANDRADE, Pedro Victor Silva de. **Tutela da honra nas redes sociais: a contribuição possível da teoria da impolidez**. 2019. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: <https://repositorio.ufmg.br/server/api/core/bitstreams/235d96e8-abe6-4e23-b609-5f5f814b47b1/content>. Acesso em: 28 out. 2023.

BARLOW, John Perry. **A Declaration of Independence of Cyberspace**. Suíça: Electronic Frontier Foundation, 1996. Disponível em: <https://www.eff.org/pt-br/cyberspace-independence>. Acesso em: 10 jul. 2024.

BEM PARANÁ. **Contra fake news, Instagram e Facebook colocam avisos em postagens sobre Eleições 2022**. Bem Paraná, Curitiba, 2021. Disponível em: <https://www.bemparana.com.br/publicacao/blogs/politicaemdebate/contra-fake-news-instagram-e-facebook-colocam-avisos-em-postagens-sobre-eleicoes-2022/>. Acesso em: 12 ago. 2024.

BRASIL. 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 abr. 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 12 ago. 2024.

CALVINO, Ítalo. **O visconde partido ao meio**. Tradução de Nilson Moulin. São Paulo: Companhia das Letras, 2009.

CASTELLS, Manuel. **A sociedade em rede**. Tradução de Roneide Venâncio Majer. São Paulo: Paz e Terra, 2002. v. 1. (A era da informação: economia, sociedade e cultura).

CASTELLS, Manuel; CARDOSO, Gustavo. **A sociedade em rede: do conhecimento à acção política**. Belém, Portugal: Imprensa Nacional, Casa da Moeda, 2005.

CHAVEZ, Nicole et al. **Pittsburgh synagogue gunman said he wanted all Jews to die, criminal complaint says**. CNN, Estados Unidos, 2018. Disponível em: <https://edition.cnn.com/2018/10/28/us/pittsburgh-synagogue-shooting/index.html>. Acesso em: 6 ago. 2024.

COMISSÃO EUROPEIA. **Call for participants: RAN C&N meeting on How to deal with borderline content (related to hate speech, meme culture, humour etc.) from the perspective of public trust? Face-to-face meeting, Rotterdam, The Netherlands on 16 and 17 April 2024**. European Commission: Migration e Home Affairs, News Announcement, 01 mar. 2024. Disponível em: https://home-affairs.ec.europa.eu/news/call-participants-ran-cn-meeting-how-deal-borderline-content-related-hate-speech-meme-culture-humour-2024-03-01_en. Acesso em: 6 ago. 2024.

CONSULTOR JURÍDICO. MPF e DPU acionam Google na Justiça por discurso de ódio de policiais. **Consultor Jurídico**, São Paulo, 4 maio 2024. Disponível em: <https://www.conjur.com.br/2024-mai-04/mpf-e-dpu-acionam-google-na-justica-por-discurso-de-odio-de-policiais/>. Acesso em: 14 ago. 2024

D'ANDREA, Carlos Frederico de Brito. **Pesquisando plataformas digitais: conceitos e métodos**. Salvador: UFBA, 2020.

FILGUEIRAS, Fernando; ALMEIDA, Virgílio. **Governance for the digital world: neither more state nor more market**. Suíça: Palgrave Macmillan, 2021.

GAO, Lei; HUANG, Ruihong. Detecting online hate speech using context aware models. **Proceedings of recent advances in natural language processing**, Varna, Bulgaria, p. 260-266, set. 2017. Disponível em: <https://www.acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP036.pdf>. Acesso em: 13 ago. 2024.

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

GOOGLE. **Relatório de Transparência**. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: <https://transparencyreport.google.com/youtubepolicy/removals>. Acesso em: 7 ago. 2024.

GOOGLE. **Relatório de Transparência**: Janeiro de 2024 a Março de 2024. Cumprimento das diretrizes da comunidade do Youtube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2024Q1&lu=total_removed_videos&content_by_flag=period:2024Q1;exclude_automated:all&videos_by_reason=period:2024Q1&channels_by_reason=period:2024Q1&comments_by_source=period:2024Q1&comments_removal_reason=period:2024Q1&videos_by_country=period:2024Q1;region:&videos_by_views=period:2024Q1&total_comments_removed=period:2024Q1&total_removed_videos=period:2024Q1;exclude_automated:all. Acesso em: 7 ago. 2024.

GUSTIN, Miracy Barbosa de Souza; DIAS, Maria Tereza Fonseca; NICÁCIO, Camila Silva. **(Re)pensando a pesquisa jurídica**: teoria e prática. Belo Horizonte: Almedina, 2020.

HIETANEN, Mika; EDDEBO, Johan. Towards a definition of hate speech with a focus on online contexts. Disponível em: **Journal of Communication Inquiry**, Londres, v. 47, n. 4, p. 441–458, out. 2022. <https://journals.sagepub.com/doi/epub/10.1177/01968599221124309>. Acesso em: 7 ago. 2024.

INSTAGRAM. **Why you're seeing a warning before you can view a photo or video on Instagram**. Instagram, 2024. Disponível em: https://help.instagram.com/188848648282410?helpref=uf_permalink. Acesso em: 12 ago. 2024.

JOROUVÁ, Vera. **Code of Conduct for Illegal online hate speech**: questions and answers. European Commission, Bruxelas, jun. 2016. Disponível em: https://ec.europa.eu/newsroom/document.cfm?doc_id=41844. Acesso em: 6 ago. 2024.

KEMP, Simon. **Digital 2024**: Brazil. Data Reportal, 23 fev. 2024. Disponível em: <https://datareportal.com/reports/digital-2024-brazil>. Acesso em: 12 jun. 2024.

KLEINWÄCHTER, Wolfgang; ALMEIDA, Virgílio A.F. The internet governance ecosystem and the rainforest. **IEEE Internet Computing**, Washington, v. 19, n. 2, p. 64–67, 2015. Disponível em: <https://dl.acm.org/doi/abs/10.1109/MIC.2015.49>. Acesso em: 12 jun. 2024.

KRANZBERG, Melvin. Technology and History: “Kranzberg’s Laws”. **Technology and Culture**, Baltimore, v. 27, n. 3, p. 544–560, 1986. Disponível em: <https://doi.org/10.2307/3105385>. Acesso em: 15 jul. 2026.

KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. **Transparência da moderação de conteúdo**: tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: <https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf>. Acesso em: 15 jul. 2026.

LESSIG, Lawrence. **Code and other laws of cyberspace**: version 2.0. Nova Iorque, EUA: Basic Books, 2006.

LUNA, Florencia. Elucidating the concept of vulnerability. **The International Journal of Feminist Approaches to Bioethics**, Toronto, v. 2, n. 1, p. 120-138, 2009.

MACKLIN, Graham. The Christchurch Attacks: livestream terror in the viral video age. **Combating Terrorism Center**, Nova Iorque, v. 12, n. 6, jul. 2019. Disponível em: <https://ctc.westpoint.edu/christchurch-attacks-livestream-terror-viral-video-age/>. Acesso em: 8 ago. 2024.

MARQUES, Claudia Lima; MUCELIN, Guilherme. Vulnerabilidade na era digital: um estudo sobre os fatores de vulnerabilidade da pessoa natural nas plataformas, a partir da dogmática do Direito do Consumidor. **Civilistica.com**, Rio de Janeiro, v. 11, n. 3, p. 1-30, dez. 2022. Disponível em: <https://civilistica.emnuvens.com.br/redc/article/view/872>. Acesso em: 7 nov. 2023.

MARTINS, Anna Clara Lehmann. Discurso de ódio em redes sociais e reconhecimento do outro: o caso M. **Revista Direito GV**, São Paulo, v. 15, n. 1, p. 1-30, jan./abr. 2019. Disponível em: <https://www.scielo.br/j/rdgv/a/WPZBfgrv6Md957dSxz7Hh5h/?format=pdf&lang=pt>. Acesso em: 7 nov. 2023.

MCCROSKY, Jesse; GEURKINK, Brandi. **Youtube Regrets**: A crowdsourced investigation into Youtube's recommendation algorithm. Mozilla Foundation, Califórnia, 2021. Disponível em: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Acesso em: 30 out. 2023.

MCGRADY, Ryan. **What we discovered on Deep Youtube**. The Atlantic, Washington, 26 jan. 2024. Disponível em: <https://www.theatlantic.com/technology/archive/2024/01/how-many-videos-youtube-research/677250/>. Acesso em: 26 maio 2025.

MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Inquérito Civil n. 1.24.000.000812/2021-0**. Paraíba, out. 2019. Disponível em: <https://uploads.polemicaparaiba.com.br/2021/06/ACP-Discurso-de-odio-contra-mulheres-na-Paraiba.pdf>. Acesso em: 14 ago. 2024.

MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República na Paraíba. **Procedimento Investigativo Criminal n. 1.24.000.000924/2021-51**. Paraíba, jan. 2023. Disponível em: <https://static.poder360.com.br/2023/02/pedido-prisao-sikerajunior-mpf.pdf>. Acesso em: 14 ago. 2024.

MINISTÉRIO PÚBLICO FEDERAL. Procuradoria da República no Rio de Janeiro. **Ação Civil Pública n. 5028178-59.2024.4.02.5101**. Rio de Janeiro, abr. 2024. Disponível em: <https://tinyurl.com/4ddzy6ec>. Acesso em: 14 ago. 2024.

MOST popular social networks worldwide as of April 2024, ranked by number of monthly active users (in millions). Statista, ago. 2024. Disponível em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 12 ago. 2024.

MUNGER, Kevin; PHILLIPS, Joseph. Right-Wing YouTube: a supply and demand perspective. **The International Journal of Press/Politics**, Califórnia, v. 27, n. 1, p. 186-219, 2022. Disponível em: <https://journals.sagepub.com/doi/10.1177/1940161220964767>. Acesso em: 21 ago. 2024.

NALON, Tai. Notas da Comunidade do ex-Twitter são capturadas por grupos e expõem usuários ao ódio. **Terra**, São Paulo, 27 out. 2023. Disponível em: <https://www.terra.com.br/noticias/checamos/notas-da-comunidade-do-ex-twitter-sao-capturadas-por-grupos-e-expoem-usuarios-ao-odio,1874f0d37413940fe513015e62f3523as5luexff.html>. Acesso em: 21 ago. 2024.

OBAR, Johnatan A.; WILDMAN, Steve. Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. **Telecommunications Policy**, Oxford, v. 39, n. 9, p. 745-750, 2015. Disponível em: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2927584_code962598.pdf?abstractid=2647377&mirid=1&type=2. Acesso em: 15 ago. 2024.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **UN strategy and plan of action on hate speech**. ONU, Nova Iorque, maio 2019. Disponível em: https://www.un.org/sites/un2.un.org/files/action_plan_on_hate_speech_en.pdf. Acesso em: 12 ago. 2024.

ORGANIZAÇÃO DOS ESTADOS AMERICANOS. Convenção Interamericana para o combate a toda forma de discriminação e intolerância. OEA, Guatemala, 05 jun. 2013. Disponível em: https://www.oas.org/en/sla/dil/docs/inter_american_treaties_A-69_Convencao_Interamericana_discriminacao_intolerancia_POR.pdf. Acesso em: 8 ago. 2024.

POLAND, Baley. **Haters: harassment, abuse and violence online**. Nebraska, EUA: Potomac Books, 2016.

RAINIE, Lee; ANDERSON, Janna; ALBRIGHT, Jonathan. **The future of free speech, trolls, anonymity and fake news online**. Pew Research Center, Washington, 29 mar. 2017. Disponível em: <https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>. Acesso em: 25 mar. 2024.

RECHMANN, Aline. **Aprovação do PL das Fake News é improvável, diz pesquisa**. Gazeta do Povo, Curitiba, 30 maio 2024. Disponível em: <https://www.gazetadopovo.com.br/república/45-dos-deputados-consideram-improvavel-a-aprovacao-do-pl-das-fake-news/>. Acesso em: 14 ago. 2024.

RIBEIRO, Manoel Horta et al. **Characterizing and Detecting Hateful Users on Twitter**. 23 mar. 2018. Disponível em: <https://arxiv.org/pdf/1803.08977>. Acesso em: 8 ago. 2024.

RIBEIRO, Manoel Horta; OTTONI, Raphael et al. Auditing Radicalization Pathways on YouTube. In: PROCEEDINGS of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: Association for Computing Machinery, 2020. (FAT* '20), p. 131–141. ISBN 9781450369367. DOI: 10.1145/3351095.3372879.

RUEDIGGER, Marco Aurélio. **Regulação de plataformas digitais: uma regulação para a análise do debate nacional frente a um desafio global**. Rio de Janeiro: FGV DAPP, 2022.

SAID, Edward. **Orientalismo: o Oriente como invenção do Ocidente**. São Paulo: Companhia das Letras, 2007.

SANDER, Barrie. Freedom of expression in the age of online platforms: the promise and pitfalls of a human-rights based approach to content moderation. **Fordham International Law Journal**, Nova Iorque, v. 43, p. 939-1006, 2019. Disponível em: <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2787&context=ilj>. Acesso em: 8 ago. 2024.

SLOANE, Garret. **Stop Hate for Profit: os efeitos do boicote um ano depois**. Meio e Mensagem, São Paulo, 2024. Disponível em: <https://www.meioemensagem.com.br/midia/stop-hate-for-profit-os-efeitos-do-boicote-um-ano-depois>. Acesso em: 8 ago. 2024.

TUFEKCI, Zeynep. Big questions for social media big data: representativeness, validity and other methodological pitfalls. **ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media**, 2014. Disponível em: <https://arxiv.org/pdf/1403.7400>. Acesso em: 12 ago. 2024.

TWITTER/X. **Freedom of Speech, Not Reach: An update on our enforcement philosophy**. X Safety, Texas, 17 abr. 2023. Disponível em: https://blog.x.com/en_us/topics/product/2023/freedom-of-speechnot-reach-an-update-on-our-enforcement-philosophy. Acesso em: 12 ago. 2024.

VALENTE, Jonas. Regulando desinformação e fake news: um panorama internacional das respostas ao problema. **Comunicação Pública**, Lisboa, v. 14, n. 27, p. 1-17, 2019. Disponível em: <https://journals.openedition.org/cp/5262#quotation>. Acesso em: 12 ago. 2024.

VIEIRA, Alessandro. **Projeto de Lei nº 2630/2020**. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. Brasília: Senado Federal, 13 maio 2020. Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&ts=1712173889308&disposition=inline>. Acesso em: 14 ago. 2024.

VIEIRA, Alessandro. **Emenda ao Projeto de Lei nº 2630/2020 - SF/20458.58654-0**. Brasília: Senado Federal, 2020. Disponível em: <https://legis.senado.leg.br/sdleg-getter/documento?dm=8110634&disposition=inline>. Acesso em: 14 ago. 2024.

VOLPINI, Silvia Pires. **Sentidos da vulnerabilidade no contexto da deficiência a partir da análise de decisões judiciais**. 2023. Dissertação (Mestrado em Direito) – Universidade Federal de Minas Gerais, Belo Horizonte, 2023. Disponível em: <https://repositorio.ufmg.br/handle/1843/55177>. Acesso em: 10 ago. 2024.

YOUTUBE. **Age-restricted content**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2802167?hl=en>. Acesso em: 12 ago. 2024.

YOUTUBE. **Hate speech policy**. San Bruno: Youtube, 2019. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=en>. Acesso em: 6 ago. 2024.

YOUTUBE. **Políticas contra spam, práticas enganosas e golpes**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2801973?hl=pt-BR>. Acesso em: 6 ago. 2024.

YOUTUBE. **Report inappropriate videos, channels, and other content on YouTube**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/2802027>. Acesso em: 6 ago. 2024.

YOUTUBE. **Youtube channel monetization policies**. San Bruno: Youtube, 2024. Disponível em: <https://support.google.com/youtube/answer/1311392?hl=en>. Acesso em: 12 ago. 2024.

YOUTUBE TEAM. **Our ongoing work to tackle hate**. San Bruno: Youtube Official Blog, maio 2019. Disponível em: <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/>. Acesso em: 6 ago. 2024.

YOUTUBE VIWERS. **The life of a flag**. San Bruno: Youtube, 2018. Disponível em: <https://www.youtube.com/watch?v=WK8qRNSmhEU>. Acesso em: 6 ago. 2024.

WYZOL. **Youtube Stats**: 2025. Disponível em: <https://www.wyzowl.com/youtube-stats/>. Acesso em: 26 maio 2025.