Revista Jurídica Cesumar
DOI: 10.17765/2176-9184.2024v24n1.e12690

A HUMANIZAÇÃO DO COMBATE À DESINFORMAÇÃO

THE HUMANIZATION OF COMBATING DISINFORMATION

LA HUMANIZACIÓN DEL COMBATE CONTRA LA DESINFORMACIÓN

Felipe Chiarello[*]
Elon Caropreso Herrera[**]
Fernando Franco Barbosa Filho[***]

RESUMO: O presente artigo apresenta e discute os efeitos deletérios da desinformação ao regime democrático brasileiro. Para isso, busca contextualizar o atual debate entorno dos limites da liberdade de expressão, conceituando as condutas criminosas no ambiente virtual e o discurso de ódio. Com a finalidade de compreender os mecanismos legais e contratuais disponíveis para coibir esse tipo de prática, o artigo explora brevemente algumas ferramentas de controle oferecidas pelas redes sociais. À luz dessas ferramentas e recursos disponíveis, propõe-se a reflexão sobre a necessidade de humanizar as tratativas pelas redes sociais quando formalizadas queixas ou denúncias de conteúdo ofensivo por parte de seus usuários.

PALAVRAS-CHAVE: desinformação; controle; redes sociais; humanização; Democracia.

ABSTRACT: This article presents and discusses the deleterious effects of misinformation on the Brazilian democratic system. To achieve this, it seeks to contextualize the current debate surrounding the limits of freedom of expression, conceptualizing criminal behaviors in the virtual environment and hate speech. In order to understand the legal and contractual mechanisms available to curb such practices, the article briefly explores some control tools offered by social media platforms. In light of these tools and resources, the article proposes a reflection on the need to humanize the procedures undertaken by social media platforms when addressing complaints or reports of offensive content submitted by their users.

KEY-WORDS: disinformation; control; social networks; humanization; Democracy.

RESUMEN: Este artículo presenta y discute los efectos deletéreos de la desinformación en el régimen democrático brasileño. Para ello, busca contextualizar el actual debate en torno a los límites de la libertad de expresión, conceptualizando las conductas criminales en el ambiente

[*] Advogado. Atual Diretor da Faculdade de Direito da Universidade Presbiteriana Mackenzie (UPM), cargo que já havia ocupado de 2016 a 2020, Professor Titular da Faculdade de Direito e do Programa de Mestrado e Doutorado em Direito Político e Econômico. Foi Pró-Reitor de Pesquisa e Pós-Graduação da UPM (2020-2023). Mestre e Doutor em Direito pela PUC/SP.

[**] Mestre em Direito Político e Econômico pela Universidade Presbiteriana Mackenzie, com bolsa integral pela CAPES. Durante a graduação, cursou extensão universitária pela Faculdade de Direito da Universidade de Coimbra (2013/2014). É secretário-geral do grupo de pesquisa Estado e Economia no Brasil. Membro-associado do Instituto Brasileiro de Estudos Constitucionais-IBEC. Sócio no escritório Dinamarco, Rossi, Beraldo & Bedaque Advocacia.

[***] Pós-graduando em Direito Civil e Direito do Consumidor pela Escola Paulista de Direito (EPD). Graduado em Direito pela Universidade Presbiteriana Mackenzie. Membro da Comissão de Direito Civil da OAB-SP. Integra o Grupo de Estudos Avançados em Processo da Faculdade de Direito da Universidade de São Paulo.

Recebido em: 12/04/2024
Aceito em: 20/04/2024

DOI: 10.17765/2176-9184.2024v24n1.e12690

virtual y el discurso de odio. Con el fin de comprender los mecanismos legales y contractuales disponibles para cohibir este tipo de práctica, el artículo explora brevemente algunas herramientas de control ofrecidas por las redes sociales. A la luz de estas herramientas y recursos disponibles, se propone la reflexión sobre la necesidad de humanizar los tratos por las redes sociales cuando se formalizan quejas o denuncias de contenido ofensivo por parte de sus usuarios.

PALABRAS-CLAVE: desinformación; control; redes sociales; humanización; Democracia.

INTRODUÇÃO

Devido às experiências adquiridas pela humanidade nos últimos três ou quatro séculos, consolidou-se no ocidente a máxima de preservação do regime democrático em detrimento às experiências absolutistas e ditatoriais já vivenciadas. O avanço tecnológico e a vastidão de conteúdo ofertado no mundo virtual impulsionaram a disseminação dessa tendência universal, modificando a maneira de relação entre as nações, tanto econômica quanto politicamente. Nesse trajeto, porém, foi subestimada a capacidade da tecnologia em se prestar à veiculação de ideias que tendam a minar os próprios valores plurais da Democracia.

Por força dessa preocupante aptidão, os Estados se viram obrigados a editar atos normativos para delimitação do lícito e ilícito no ambiente virtual. Mais do que isso, a fim de se evitar a prática de atos criminosos sob o manto do anonimato digital, as estruturas estatais tiveram de sofisticar os seus meios de controle, passando a contar com a atuação colaborativa das plataformas digitais.

Em decorrência das disputas judiciais envolvendo os usuários e os provedores de internet ou de aplicação, ressurgiram profundos e relevantes debates filosófico-jurídicos, tais como aquele que envolve os limites da liberdade de expressão, sobretudo se considerada a amplitude de alcance das redes sociais. Desinformação, *fake news* e discurso de ódio são alguns dos elementos com os quais o Poder Judiciário e a sociedade civil devem aprender a combater. Nesta batalha pela prevalência democrática, entretanto, não pode ser absorto o comportamento das plataformas digitais, cumprindo a elas um papel ativo de construção jurídica, por meio de uma atuação cooperativa e responsiva.

Afinal, a judicialização dessas questões não parece ter atendido a contento a demanda dos usuários por cessar condutas criminosas. Sob uma ótica mais ampla de observação, e considerando o Poder Público como verdadeiro elaborador e catalizador de políticas públicas, talvez incumba a ele uma atuação mais pontual e incisiva na regulamentação da atividade de moderação das redes sociais, exigindo uma conduta mais especializada e humanizada dessas plataformas, a fim de se garantir à rede mundial de computadores um ambiente sadio e mais equilibrado.

1 O ATUAL ESTÁGIO DO DEBATE SOBRE OS LIMITES DA LIBERDADE DE EXPRESSÃO: CONDUTAS CRIMINOSAS E DISCURSO DE ÓDIO

No Brasil, a regra da liberdade de expressão (art. 5º, inc. IV)[1] foi erigida à condição de norma constitucional de *eficácia plena*[2] e aplicabilidade imediata (§ 1º do art. 5º) pelo constituinte originário. Isso significa dizer que sua grandeza, enclausurada no extenso rol de garantias disposto pelo art. 5º da Constituição da República, impõe-se como regra de observância geral pela sociedade. A condicionante incluída pela parte final do inc. IV,[3] porém, já prenuncia o debate que ora se propõe aprofundar. Porquanto o nascedouro (ou ressurreição, a depender do ponto de vista) do regime democrático brasileiro esteja vinculado ao fim do funesto episódio autoritário, é conhecido o esforço da Assembleia Nacional Constituinte, liderada pelo Deputado Ulysses Guimarães, em eliminar quaisquer resquícios característicos do período antidemocrático vivenciado, a exemplo da conhecida e reiterada prática de censura durante os *anos de chumbo*. Desde então, a liberdade de expressão simboliza a consolidação do regime

[1] "Art. 5º Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes: (...) IV – é livre a manifestação do pensamento, sendo vedado o anonimato; (...)".

[2] "As normas de eficácia plena incidem diretamente sobre os interesses a que o constituinte quis dar expressão normativa. São de aplicabilidade imediata, porque dotadas de todos os meios e elementos necessários à sua executoriedade. No dizer clássico, são autoaplicáveis. As condições gerais para essa aplicabilidade são a existência apenas do aparato jurisdicional, o que significa: aplicam-se só pelo fato de serem normas jurídicas, que pressupõem, no caso, a existência do Estado e de seus órgãos" (SILVA, José Afonso da. *Aplicabilidade das normas constitucionais*. 8ª ed. São Paulo: Malheiros, 2012, p. 100).

[3] "(...), sendo vedado o anonimato".

democrático de direito.[4] Apesar disso, é premissa para o exercício desse direito – liberdade de manifestação – a plena possibilidade de identificação uma vez externalizada a expressão. Daí diz-se que vigora, assim, o regime da liberdade com responsabilidade.

Sob o imenso guarda-chuva da dignidade da pessoa humana, capitaneado pelo Supremo Tribunal Federal, o Judiciário tenta aplicar a conhecida técnica de ponderação de bens,[5] com a finalidade de encontrar um ponto de equilíbrio entre eventuais valores fundamentais em confronto com a liberdade de expressão. Faltam, no entanto, objetividade na análise jurisprudencial e parâmetros concretos para, caso a caso, deliberar sobre a extensão desse importante direito fundamental de primeira dimensão.[6]

Talvez o mais emblemático episódio apreciado pelo Supremo Tribunal Federal seja, de fato, o caso Ellwanger. Nele, Sigfried Ellwanger foi processado e julgado por publicar, vender e disseminar conteúdo antissemita, tipificando-se sua conduta como crime de racismo. O escritor negacionista fundou sua própria editora para, sobretudo, propagar ideais negacionistas da existência do holocausto. Defronte ao *habeas corpus* impetrado pela defesa do gaúcho (HC 82424-2), aquela Corte entendeu que a falsa premissa biológica da qual partira o paciente – existência de raças como espécie de subdivisão da condição humana – consistir-se-ia em berço do racismo e da ideia de discriminação e segregação racial. Sobre o conteúdo decisório e os debates daquela Corte, Ronaldo Porto Macedo Junior observou que, a despeito de recorrerem ao mesmo marco teórico para fins de ponderação (R. Alexy), Gilmar Mendes e Marco Aurélio Mello expuseram conclusões diametralmente opostas:

> É interessante notar que o mesmo sentimento da falta de padrões claros e tensão entre diferentes conceituações de liberdade expressão também é encontrado no caso Ellwanger, muitas vezes tratado como o principal caso sobre o tema. É particularmente revelador que os Ministros Marco Aurélio Mello e Gilmar Mendes Ferreira tenham invocado a doutrina de ponderação de Alexy, mas tenham alcançado opiniões contrárias. É ainda mais surpreendente como ambos os ministros depois de fazerem um breve resumo acerca do teste de proporcionalidade simplesmente pulam para a conclusão sem maiores justificativas (tradução livre).[7]

Este breve excerto demonstra que é antiga a complexidade das soluções adotadas pelo Supremo Tribunal Federal em matéria de liberdade de expressão. É justamente por essa razão que, ao contextualizar esse debate, mostra-se fundamental conhecer algumas outras experiências estrangeiras, as quais talvez permitam ao interlocutor identificar as tendências do Poder Judiciário brasileiro para o enfrentamento de temas dessa envergadura.

Para facilitar o exercício a que se propõe este artigo, vale a menção ao estudo feito por Geoffrey R. Stone acerca das dez lições tidas durante o século XX pela Suprema Corte dos Estados Unidos no enfrentamento de casos envolvendo o conteúdo da Primeira Emenda.[8]

---

[4] FILHO, Adalberto Camargo Aranha; PINTO, Felipe Chiarello de Souza; RAMOS, Tais. *A Liberdade de Expressão Política e o Discurso de Ódio (HateSpeech) em Período Eleitoral: estudos de casos da Corte Interamericana de Direitos Humanos*. *In*: MARQUES, Claudia Lima; MARTINI, Sandra Regina; FINCO, Matteo (org). *Diálogos entre direitos humanos, direito do consumidor, compliance e combate à corrupção*. 1. ed. São Paulo: YKEditora, 2021, p. 155.

[5] A complexa técnica consagrada na Alemanha, resultante da pesquisa de Robert Alexy, consiste em ser capaz de identificar qual princípio "*tem precedência em face do outro sob determinadas condições*", de modo que os "*conflitos entre regras ocorrem na dimensão da validade, enquanto as colisões entre princípios – visto que só princípios válidos podem colidir – ocorrem, para além dessa dimensão, na dimensão do peso*" (*in* ALEXY, Robert. *Teoria dos direitos fundamentais*. 2ª ed. 4ª tiragem. São Paulo: Malheiros, 2015, p. 94 e ss.).

[6] - Não se desconhece a divergência doutrinária acerca da utilização da designação de *dimensão* ou *geração* para as diferentes correntes de direitos fundamentais experimentadas pela humanidade. Aliando-se à utilização da nomenclatura *geração*, estão os juristas Paulo Bonavides, Manoel Gonçalves Ferreira Filho, Gilmar Mendes e Paulo Branco, Alexandre de Moraes, entre tantos outros. Em contraponto à ideia de conflito geracional ou mesmo substituição das gerações, capitaneiam a ideia de *dimensões* os especialistas Ingo Sarlet, Cançado Trindade, José Joaquim Gomes Canotilho, entre tantos outros. A adoção de uma ou outra designação não afasta o propósito da utilização dessa referência: socorrer-se dos direitos fundamentais atrelados às liberdades do sujeito.

[7] No original: "It is worth noticing that the same feeling of lack of clear standards and tension among different conceptualizations of free speech are also found in the Ellwanger case, often treated as a leading case on the matter. It is particularly revealing that Justices Marco Aurelio Mello and Gilmar Mendes Ferreira invoked Alexy's doctrine of balancing but arrived at contrary opinions! It is quite surprising how both justices after making a summary of the proportionality test simply jumped to a conclusion without further justification" (*in* MACEDO JUNIOR, Ronaldo Porto Macedo. *Freedom of Expression: what lessons should we learn from US experience?* São Paulo: Revista Direito GV, v. 13, n. 1, págs. 274-302, jan-abr/2017. Disponível em <https://www.scielo.br/j/rdgv/a/tRnqx97GRkqny4L77JFGBTx/?format=pdf&lang=en>). Acesso em 6 jun. 2023.

[8] STONE, Geoffrey R. *Free Speech in the T ee Speech in the Twenty-First Centur wenty-First Century: Ten Lessons fr en Lessons from the Twentieth Century*. Rev. Pepperdine Law Review, v. 36, Issue 2, mai-2009, págs. 273-300. Disponível em <https://digitalcommons.pepperdine.edu/cgi/viewcontent.cgi?article=1072&context=plr >. Acesso em 6 jun. 2023.

Ao se aprofundar nas lições apresentadas por Stone, Porto Macedo avalia que há três pontos na jurisprudência americana acerca do *freedom of speech* que são dignos de nota.

Em primeiro lugar, o fato de que, (a) nos Estados Unidos, a doutrina da Primeira Emenda levou mais a sério a abordagem filosófica e política envolvendo a definição e extensão da liberdade de expressão. Para tanto, foram traçadas justificativas políticas de existência da livre manifestação, entre as quais se destacam (a.1) o autogoverno, que é pressuposto dessa discussão sob a perspectiva da lógica democrática; (a.2) a mídia, também chamada de quarto poder; (a.3) um mercado competitivo de ideias, em que usualmente a verdade figure como elemento mais disputado e concorrido; e (a.4) a autodeterminação e a autonomia individual dos cidadãos, que *a priori* garantiriam o discurso não-político, mas *a posteriori* não seriam as únicas alternativas de resguardar a autonomia e autodeterminação dos indivíduos.

Em segundo lugar, ainda de acordo com o autor e, a nosso juízo, como consequência natural do primeiro ponto, há (b) o desenvolvimento de um aparato conceitual sólido para resolução de questões envolvendo a Primeira Emenda, o que viabiliza a melhor solução para cada caso, com base nesse aparato sólido e sem recorrer a apressadas hipóteses exageradas ou absurdas para testar e flexionar argumentos.[9]

Por fim, o último ponto apresentado está atrelado (c) à rejeição de três abordagens nocivas de interpretação da Primeira Emenda, sendo elas (c.1) a absolutista, que poderia ser traduzida à interpretação literal; (c.2) a ideia de que somente seria cabível um único padrão de revisão para todos os casos envolvendo a Primeira Emenda, de tal maneira a garantir que ora se dê maior proteção à liberdade de expressão ora aos interesses estatais; e, ainda, (c.3) a parcial, cautelosa e sensível adesão à tese de balanceamento *ad hoc*, que garante justamente o afastamento da incerteza e imprevisibilidade nos julgamentos.[10]

Em seu conhecido *Quatro ensaios sobre a liberdade*, Isaiah Berlin contrapõe duas perspectivas opostas para a concepção de liberdade, sendo a primeira delas marcada pelo seu âmbito negativo, que se caracteriza pela inoponibilidade de ingerências naquilo que se o cidadão diz, e seu âmbito positivo, marcado pela liberdade de participar politicamente das decisões sociais (*freedom of* e *freedom to*).[11] No caso americano, é comum encontrar casos em que a jurisprudência tem assegurado a proteção integral e irrestrita a ambas essas esferas de liberdade. Sobre isso, vale observar os dois precedentes analisados por Riva Sobrado de Freitas e Matheus Felipe de Castro.

> (...) Decisão importante, reveladora dessa proteção, está referida no caso Brandenburg *versus* Ohio em 1969, no qual um indivíduo líder da Ku Klux Klan foi condenado pela Suprema Corte de Ohio por fazer apologia ao crime, defender a violência e os métodos de terrorismo como meio de empreender reforma política e industrial.
>
> Em uma reunião filmada por um jornalista convidado havia ainda o pronunciamento de palavras de ordem valorizando brancos e depreciando negros e judeus. Essa decisão foi revertida pela Suprema Corte estadunidense. Os juízes Black e Douglas entenderam que a doutrina do perigo claro e eminente de dano (*clear and present danger*) não deveria ser levada em conta em tempos de paz, na interpretação da primeira emenda da Constituição. A Corte constatou, ainda, a falta de evidência do incitamento, porque a Ku Klux Klan defendeu ideias abstratas e com convicção, não tendo o governo o poder de invadir o espaço da crença.
>
> Outra decisão no sentido da proteção do *hate speech* aparece descrita no caso R.A.V. *versus* City of Saint Paul (1992), Minnesota. Nessa oportunidade, alguns adolescentes foram presos por invadir o quintal de uma família afrodescendente e atear fogo a uma cruz. A Suprema Corte de Minnesota, com base em legislação estadual, que tipificava crimes motivados por preconceito, entendeu que tal ato consistia em clara demonstração de depreciação em razão de raça e proferiu a condenação.
>
> A Suprema Corte estadunidense, entretanto, reverteu também essa decisão considerando, inclusive, inconstitucional a Lei do Estado de Minnesota, porque ela estabelecia restrições para preconceito, envolvendo raça, cor, credo religioso, etc.; proibindo palavras de ordem que contivessem o discurso do ódio. Argumentou

[9]

[10] MACEDO JUNIOR, Ronaldo Porto Macedo. *Freedom of Expression: what lessons should we learn from US experience?* São Paulo: Revista Direito GV, v. 13, n. 1, págs. 274-302, jan-abr/2017. Disponível em <https://www.scielo.br/j/rdgv/a/tRnqx97GRkqny4L77JFGBTx/?format=pdf&lang=en>. Acesso em 6 jun. 2023.

[11] BERLIN, Isaiah. *Quatro ensaios sobre a liberdade*. Trad. Wumberto Hudson Ferreira. Brasília, DF: Ed. UnB, 1981.

> ainda que o governo não pode regular categorias de discurso com base em hostilidades ou favoritismos, como os discursos depreciativos em razão de raça, por exemplo. Essas manifestações estariam protegidas pela Liberdade de Expressão e pela primeira emenda da Constituição estadunidense.[12]

A contraposição à doutrina americana baseada na interpretação da Primeira Emenda talvez resida do outro lado do globo, num país extremamente ressentido cuja liberdade de expressão excessiva e irrestrita resultou em um dos maiores genocídios da história da humanidade: a Alemanha. É compreensível – para não se dizer lógica – a preocupação do legislador alemão para com a necessidade de estabelecer limites claros para a liberdade de expressão. Exemplo de uma dessas restrições impostas pela Alemanha está na impossibilidade de se negar a existência do holocausto. Esta não é uma manifestação livre de pensamento, segundo a jurisprudência, mas a negação pura e simples de determinado fato ocorrido. Essa é a interpretação desde 1994, quando decidido pelo Tribunal Constitucional Federal Alemão o caso do historiador David Irving.[13]

Talvez o mais citado dos casos envolvendo a temática de liberdade de expressão pela jurisprudência alemã esteja ligado ao precedente do cidadão alemão Erich Lüth, um famoso produtor de cinema à época. No início da década de 50, Lüth teria convocado todos os distribuidores de filmes cinematográficos a boicotarem a película cinematográfica lançada por Veit Harlan, tendo em vista o histórico antissemita de Harlan em outros filmes anteriores, tais como "Jud Süß", em 1941. Por conta dessa tentativa de boicote, Harlan e seus parceiros ajuizaram um processo em face de Lüth, a fim de que ele fosse obrigado a cessar sua conduta convocatória, sob pena de multa. Inicialmente julgada procedente pelo Tribunal de Hamburgo, o caso veio a ser reformado pelo Tribunal Constitucional Federal Alemão, que decidiu pelo regular exercício de um direito público subjetivo de resistência vinculado à liberdade de expressão. Esse precedente teve relevantíssima figura na jurisprudência alemão, sobretudo pela criação da figura de *Drittwirkung* e *Ausstrahlungswirkung*, eficácia horizontal dos direitos fundamentais, e do efeito limitador dos direitos fundamentais contra seus próprios limites (*Wechselwirkung*), tornando-se necessária a ponderação no caso concreto.[14]

Com base nesse panorama do atual estágio do debate envolvendo a compreensão da liberdade de expressão nos Estados Unidos da América e na Alemanha – esses dois países que têm contexto histórico bastante diverso e, além disso, representam tendências às vezes opostas na compreensão jurisprudencial –, é preciso compreender a extensão da definição de discurso de ódio, que é tão fortemente combatido nos tempos atuais, bem como quais as condutas criminosas que devem ser afastadas e bloqueadas por todos agentes imersos no ambiente virtual.

Em sintética definição, e com base em relevantes autoridades teóricas sobre o assunto,[15] Riva Sobrado de Freitas e Matheus Felipe de Castro oferecem uma curta, porém proveitosa, elucidação daquilo que qualificam como discurso de ódio:

> Na busca de um conceito operacional para o discurso de ódio (*hate speech*), observa-se que tal discurso apresenta como elemento central a expressão do pensamento que desqualifica, humilha e inferioriza indivíduos e grupos sociais. Esse discurso tem por objetivo propagar a discriminação desrespeitosa para com todo aquele que possa ser considerado "diferente", quer em razão de sua etnia, sua opção sexual, sua condição econômica ou seu gênero, para promover a sua exclusão social. (...)
> (...) é possível observar que tal discriminação indica não apenas uma diferença, mas uma assimetria entre

[12] FREITAS, Riva Sobrado de; CASTRO, Matheus Felipe de. *Liberdade de Expressão e Discurso de Ódio: um exame sobre as possíveis limitações à liberdade de expressão*. Rev. Seqüência (Florianópolis), 66, p. 327-355, jul.2013. Disponível em <https://doi.org/10.5007/2177-7055.2013v34n66p327>. Acesso em 5 jul. 2023.

[13] Este célebre precedente jurisprudencial deu ensejo à criação do filme "*Denial*", do diretor inglês Mick Jackson.

[14] Para um detalhamento completo do caso, com tradução das razões e fundamentos, vale a leitura de WOISCHNIK, Jan. (*ed.*); MARTINS, Leonardo (*org.*). *Cinquenta anos de jurisprudência do Tribunal Constitucional Federal Alemão*. Trad. Beatriz Hennig, Leonardo Martins, Mariana Bigelli de Carvalho, Tereza Maria de Castro e Vivianne Geraldes Ferreira. Konrad-Adenauer-Stiftung E.V., 2005, p. 379-494.

[15] WALDRON, Jeremy. *Dignity and defamation: the visibility of hate*. Harvard Law Review, v. 123, n. 1.596, p. 1.597-1.657, 2010. LEAL DA SILVA, Rosane *et al*. *Discursos de ódio em redes sociais*: jurisprudência brasileira. Revista FGV Law, São Paulo, v. 7, n. 2, p. 445-468, jul.-dez.2011.

> duas posições: uma supostamente superior, daquele que expressa o ódio, e outra inferior, daquele contra o qual a rejeição é dirigida. O objetivo pretendido é humilhar para amedrontar pessoas ou grupos sociais evidenciando que, por suas características específicas, eles não são dignos da mesma participação política (WALDRON, 2010). Calar, excluir e alijar são propósitos da manifestação do ódio.[16]

De acordo com John S. Mill,[17] uma das finalidades da liberdade de expressão é precisamente a busca pela verdade. Isso importa dizer que, num livre mercado de ideias, a competição resultaria da disputa pela verdade, mediante o aberto debate e a vedação à censura. O problema é que, ao longo do processo de amadurecimento democrático, se percebeu ser necessária alguma limitação à liberdade de expressão, para que, paradoxalmente, pudesse ser garantido justamente o próprio direito à existência, à fala ou mesmo à reunião. Não basta, portanto, a lógica de abstenção estatal celebrada com o plexo de garantias vinculadas à primeira dimensão de direitos fundamentais. Com esse amadurecimento, percebeu-se que a manutenção do regime das liberdades, em alguma medida, clamava pela restrição a certos comportamentos e condutas justificadas anteriormente com base na liberdade de expressão. Uma dessas condutas vedadas na dinâmica democrática é o discurso de ódio, definido acima como a fala desqualificadora e humilhante, que tem por único objetivo o estabelecimento de *fator de discrímen* injusto e ilegal, incompatível com a ordem constitucional.[18]

Aqui, vale um registro democrático das opiniões científicas divergentes que, na contramão da premissa teórica adotada por esse artigo, entendem por inadequadas as restrições à liberdade de expressão a pretexto de combate do discurso de ódio, sobretudo por compreenderem que haveria uma violação explícita da autonomia individual dos cidadãos e, até mesmo, uma indevida posição paternal e interventora do Estado.[19]

À luz de todas essas relevantes considerações, o que resta a investigar adiante é se, ainda que imersa em ambiente virtual, a garantia de cumprimento à lei remanesce exclusivamente ao Estado, ou então o combate ao discurso de ódio coloca *pari passu* em colaboração todos os cidadãos e agentes privados envolvidos. É preciso compreender se a tarefa de mitigação de danos à Democracia é incumbência exclusiva do Poder Público ou se pode ser incumbência colaborativa dos cidadãos e plataformas virtuais.

2 MITIGANDO DANOS À DEMOCRACIA PELA ATUAÇÃO DAS REDES SOCIAIS

É inafastável o compromisso assumido pelas plataformas digitais e redes sociais em relação à manutenção do regime democrático. Não fosse apenas a relevância incomensurável desses provedores de aplicação como verdadeiros canais informativos dos cidadãos, o que talvez os condicione a, analogicamente e sempre quando possível, atender à finalidade informativa exigida dos programas de rádio e televisão (art. 221, inc. I da Constituição Federal), fato é que a própria subsunção da relação contratual havida entre usuário e os provedores de aplicação ao microssistema de defesa do consumidor denota a vulnerabilidade técnica, financeira e jurídica do cidadão em mitigar danos à democracia, cabendo essa responsabilidade, por consectário lógico, às plataformas digitais, que têm mais condições de prover o necessário para o combate ao discurso de ódio, à desinformação e demais fenômenos corrosivos do regime democrático.

[16] FREITAS, Riva Sobrado de; CASTRO, Matheus Felipe de. *Liberdade de Expressão e Discurso de Ódio: um exame sobre as possíveis limitações à liberdade de expressão*. Rev. Seqüência (Florianópolis), 66, p. 327-355, jul.2013. Disponível em <https://doi.org/10.5007/2177-7055.2013v34n66p327>. Acesso em 13 ago. 2023.

[17] MILL, John Stuart. *On Liberty* (1859). Kitchener: Batoche Books, 2001.

[18] Sobre os *fatores de discrímen* que preservam a isonomia e se compatibilizam com a ordem constitucional, vale a leitura de MELLO, Celso Antônio Bandeira de. *Conteúdo jurídico do princípio da igualdade*. 3ª. ed. São Paulo: Malheiros, 1999, p. 41-42.

[19] SARMENTO, Daniel. *A liberdade de expressão e o problema do hate speech*. Revista de Direito do Estado, Rio de Janeiro, v. 1, n. 4, p. 53-105, out.-dez. 2006. Disponível em: <www.dsarmento.adv.br/content/3-publicacoes/18-a-liberdade-de-expressao-e-oproblema-do-hate-speech/a-liberdade-de-expressao-e-o-problema-do-hate-speech-daniel-sarmento.pdf>. Acesso em 17 ago. 2023; BINENBOJM, Gustavo; PEREIRA NETO, Caio Mário da Silva. Prefácio. *In*: FISS, Owen M. *A Ironia da Liberdade de Expressão: Estado, Regulação e Diversidade na Esfera Pública*. Rio de Janeiro: Renovar, 2005. p. 6-7.

Antes de examinar propriamente "a quem" solidariamente incumbe a responsabilidade por mitigar danos à democracia, é preciso compreender "como" são mitigados tais danos. Ou seja, é necessário abordar o controle judicial e a dinâmica da política legislativa adotada. Pela forma instituída pelo legislador, as redes sociais, em regra, não teriam responsabilidade sobre o conteúdo postado em suas plataformas por terceiros. A exceção é resguardada na hipótese de o Poder Judiciário ter determinado a remoção de tal conteúdo e, ainda assim, os provedores de aplicação permanecerem inertes. É o que se extrai da interpretação da Seção III do Marco Civil da Internet (lei n.º 12.965, de 23 de abril de 2014). Essa regra de isenção de responsabilidade, porém, ainda não está completamente definida, sobretudo diante do Tema n.º 987, decorrente da afetação do RE 1.037.396-SP, perante o Supremo Tribunal Federal. Neste recurso extraordinário, em que já reconhecida a repercussão geral da matéria, se examinará a constitucionalidade do art. 19 da lei n.º 12.965/2014 (Marco Civil da Internet), que determina a necessidade de prévia e específica ordem judicial de exclusão de conteúdo para a responsabilização civil de provedor de *internet*, *websites* e gestores de aplicativos de redes sociais por danos decorrentes de atos ilícitos praticados por terceiros.

Em síntese, no litígio que está à base deste Tema encontra-se processo judicial, proveniente do Juizado Especial Cível de São Paulo, ajuizado por uma mulher em face do *Facebook* (posteriormente designado por *Meta*), por meio do qual a autora busca a remoção de um perfil falso criado com seus dados e que, de maneira indevida, vinha difamando seus vizinhos na cidade de Capivari, interior de São Paulo. Abstratamente, mas também com base no que se demonstrou até aqui, discute-se a necessidade de extirpação do discurso de ódio do âmbito das plataformas digitais. Essa pequena disputa judicial reflete uma infinidade de outros processos com a mesma temática, tanto é que apenas o Tema n.º 987 congrega mais de uma dezena de recursos com a mesma discussão.[20]

Se abstratamente a discussão envolve a necessidade – ou não – da remoção de conteúdo difamatório das plataformas digitais, evidentemente se discute o papel das redes sociais enquanto guardiãs da honra e moral alheias, se isso é ou não compatível com o Estado Democrático de Direito.

Se no mundo real a responsabilização por atos difamatórios incumbe àquele que profere conteúdo ofensivo, no mundo virtual a tarefa passa a ser mais complexa, pois não envolve apenas o prolator da sentença ofensora, mas também o sítio virtual em que aquela informação foi apostada, bem como sua repercussão no mundo real. Uma vez postado o conteúdo, fato é que aquela *URL*,[21] em regra, fica disponível por tempo indeterminado, a menos que seja retirado o conteúdo pelo provedor de aplicação responsável pela plataforma ou, ainda, que aquela determinada rede social deixe de existir. Nessa última hipótese, mesmo a impossibilidade de acesso a essa informação é questionável. Portanto, diferentemente do que se passa no mundo real, em que a palavra externalizada não tem a capacidade de sobreviver uma vez esgotado o recurso vocal que a verbalizou, fato é que a palavra virtual perdura e, por isso, merece uma atenção destacada do Poder Público, sobretudo em regulamentar a atuação dessas redes sociais, daí a iniciativa do Marco Civil e a tentativa do art. 19 em estabelecer um critério objetivo para responsabilização dos provedores de aplicação: o descumprimento de determinada ordem judicial.

No entanto, esse caminho de responsabilização exclusiva na hipótese de descumprimento de ordem judicial é realmente o meio mais eficaz para o combate às mazelas corrosivas do regime democrático? Talvez não seja o caminho mais eficaz, mas certamente é um dos mais aceitos pela jurisprudência. A responsabilização dos provedores, segundo Carlos Affonso Souza e Ronaldo Lemos, pode ser dividida em

> (...) três entendimentos que têm sido prevalentes na jurisprudência nacional sobre a responsabilidade civil dos provedores de aplicações de Internet: (i) a sua não responsabilização pelas condutas de seus usuários; (ii) a aplicação da responsabilidade civil objetiva, ora fundada no conceito de risco da atividade desenvolvida, ora no defeito da prestação do serviço; e (iii) a responsabilidade de natureza subjetiva, aqui também

[20] ARE 1.095.329; ARE 1.152.783; ARE 1.227.866; ARE 1.231.478; 1.241.097; ARE 1.250.684; ARE 1.259.682; ARE 1.293.761; ARE 1.343.094; ARE 1.345.125; ARE 1.386.016; RE 1.037.396; e RE 1.332.049.

[21] *Uniform Resource Locator* ou, em tradução livre, localizador uniforme de recursos. É essa a forma universalmente adotada para representação de diferentes documentos, mídia e serviços de rede na internet, capaz de fornecer a cada documento uma localização única.

> encontrando-se distinções entre aqueles que consideram a responsabilização decorrente da não retirada de conteúdo reputado como lesivo após o provedor tomar ciência do mesmo (usualmente através de notificação da vítima) e os que se entendem ser provedor responsável apenas em caso de não cumprimento de decisão judicial ordenando a retirada do material ofensivo.[22]

Independentemente da dinâmica de responsabilização que será atribuída ao provedor de aplicação, mesmo porque tal ponto ainda se encontra *sub judice* no momento de redação destas breves reflexões, fato é que o "como" se operacionaliza atualmente a mitigação de danos ao regime democrático – até por expressa escolha de política legislativa – passa quase sempre pelo Poder Judiciário por meio do controle judicial daquilo que se encontra no mundo virtual, o que acaba por tornar inócuas as disposições contratuais adesivas elaboradas pelas plataformas digitais.

É conhecida a iniciativa do grupo *Meta* em criar o Comitê de Supervisão, como espécie de órgão avaliador independente de conteúdos postados. Iniciativas dessa natureza parecem positivas, sobretudo porque desafogam o Poder Judiciário da função de arbitrar conflitos que, *a priori*, são simples à luz das disposições contratuais estabelecidas entre as partes, mas *a posteriori* podem trazer consigo deliberações complexas sobre censura prévia ou liberdade de expressão. Então, o que precisa ficar claro, como visto nos itens anteriores, é que as disposições contratuais previamente estabelecidas pelas plataformas, em regra, devem levar em consideração essas grandes diretrizes de combate à desinformação, ao discurso de ódio e à incitação criminosa, de modo que essas deliberações complexas acerca da censura ou liberdade de expressão já foram consideradas quando elaborado esse plexo de regras – ou ao menos deveriam tê-lo sido –, de modo que remanesceria ao julgador de tais conflitos, judicial ou extrajudicialmente, apenas o exame das pontuais e simples questões envolvendo o cumprimento ou não das disposições contratuais aderidas pelo usuário.

Nesse sentido, a conclusão óbvia é pela melhoria nos serviços extrajudiciais de controle de denúncias recebidas nas redes sociais.

Hoje, embora não se negue a existência de setores humanizados dentro dessas grandes empresas de tecnologia, fato é que o tratamento das denúncias passa previamente por uma espécie de burocracia virtual de apreciação da reclamação formalizada. Uma verdadeira URA[23] virtualizada, em que o usuário dificilmente tem por atendidos seus anseios e, mais do que isso, tem dificuldade em manusear e lidar com as respostas já previamente apresentadas pelo sistema *online*. Recorrendo a uma simples analogia, para que se compreenda a extensão desse problema, bastaria imaginar o mundo real hoje sendo privado do acesso às centrais de atendimento de operadoras de telefonia e internet, por exemplo. Se o canal entre consumidor e operadora fosse restrito à *URA* telefônica, qual seria o índice de satisfação dos usuários desses serviços? Qual poderia ser o grau de atendimento das demandas formalizadas por consumidores?

Especulativo, claro. No entanto, funciona metaforicamente bem para avaliar, na atualidade, a importância que deve ser dada ao atendimento humanizado na resolução de demandas. E mais. Quanto maior a humanidade e mais especialidade houver no tratamento, as soluções são mais eficazes. É o *case* de sucesso, por exemplo, da criação da plataforma *consumidor.gov.br* pela Secretaria Nacional do Consumidor (SENACON), que tem índice de solução superior a 80% (oitenta por cento) das reclamações registradas, num prazo médio de 7 (sete) dias.

3 A POSSÍVEL HUMANIZAÇÃO DO TRATAMENTO DE DENÚNCIAS E QUEIXAS NAS PLATAFORMAS DIGITAIS

O modelo atual de tratamento das denúncias oferecidas pelos usuários dos provedores de aplicação mostrou-se insuficiente para que a moderação de conteúdo *online* fosse satisfatória. Uma resposta mais eficaz consistiria

[22] SOUZA, Carlos Affonso; LEMOS, Ronaldo. *Marco civil da internet: construção e aplicação*. Juiz de Fora: Editar Editora Associada Ltda., 2016, p. 69-70.

[23] *Unidade de Resposta Audível*. Tecnologia que permite que os clientes interajam com os sistemas de atendimento das empresas por meio de uma ligação, teclas e reconhecimento de voz.

na revisão dos procedimentos internos dos provedores de aplicação, a fim de que a tramitação das solicitações de remoção de conteúdo ocorresse mediante interação humana direta. Dada a subjetividade do teor das publicações, que podem gerar distintas interpretações, em substituição à reiterada utilização de algoritmos e de respostas automatizadas, a tramitação das solicitações de remoção de conteúdo por meio da comunicação entre o usuário e o moderador representaria uma enorme evolução. Havendo uma interação real durante a tramitação da denúncia de determinada postagem, tanto o denunciante como o denunciado teriam melhores condições de compreender as razões pelas quais o conteúdo foi classificado ou legal ou ilegal.

Em determinados casos, é simplificada a identificação de postagens que violam diretrizes de plataformas, como nas situações em que o conteúdo discute temas categoricamente proibidos, como a pornografia infantil. Os padrões utilizados para tal remoção são concretos, eliminando margens para interpretações subjetivas. Todavia, nem todas as infrações aos termos de uso se apresentam com tamanha evidência, como ocorre em postagens que, à primeira vista, parecem meras expressões de posicionamento político, mas, ao se realizar uma análise meticulosa, revelam-se carregadas de um teor antidemocrático. Em tais circunstâncias, a questão demanda uma apreciação mais aprofundada e humanizada para se estabelecer se o conteúdo de fato transgride as normas de uso estabelecidas.

Não se pode ignorar que a inteligência artificial assume uma posição preponderante na tarefa de moderar e eliminar conteúdos nocivos na esfera digital. Em virtude do crescimento exponencial do volume de conteúdo produzido nas redes sociais, a exigência por uma moderação eficiente se tornou mais crucial do que jamais se viu. Nesse cenário, os moderadores humanos de conteúdo são confrontados com uma série de desafios, dentre os quais se destacam o aumento contínuo de conteúdo prejudicial e o impacto psíquico de moderar tais conteúdos. No ano de 2019, foi publicizado que os moderadores do *Facebook* correm o risco de desenvolver Transtorno de Estresse Pós-Traumático (TEPT), como consequência da exposição repetida a tais conteúdos perturbadores.[24]

Tal cenário contribui para uma maior utilização da inteligência artificial com o objetivo de administrar o volume crescente de conteúdo danoso, limitando, assim, a exposição humana a ele. Empresas de tecnologia, a exemplo da *Google*, estão se dedicando ao desenvolvimento de sistemas de inteligência artificial com o propósito de detectar e alertar sobre comentários de natureza tóxica e inflamatória. Tais sistemas, como o *Conversation AI*, uma iniciativa do *Google's Jigsaw*, empregam algoritmos com o intuito de identificar a linguagem nociva e sinalizá-la para posterior revisão ou exclusão.[25]

A inteligência artificial é o resultado de uma interação entre a lógica filosófica, matemática e computacional: a filosofia organiza o pensamento em estruturas, a matemática fornece os elementos necessários para a automação, enquanto a probabilidade auxilia na tomada de decisão e a computação, especialmente pela capacidade de processamento, possibilita a execução dessas operações em uma velocidade superior a dos humanos.[26] Entretanto, por mais sofisticados que sejam, esses mecanismos de inteligência artificial não se encontram imunes a falhas:

> O ser humano aprende por experiência, incluindo, nesse aspecto, a repetição. (...) A inteligência artificial não é diferente e, também, aprende por repetição e correção de rota. Ao entender um sistema híbrido que apresentará decisões e será calibrado por um humano sempre que houver um erro, ocorre dois fenômenos interessantes de ser estudado. Ao mesmo tempo que essa calibração subordina a inteligência artificial à supremacia humana e garante a presença, ela também levanta a questão do erro. Isso significa dizer que eles acontecerão, mas qual seria sua dimensão?[27]

[24] NEWTON, Casey. *Bodies in seats.* he Verge: Vox Media, Estados Unidos, 19 de junho de 2019. Disponível em: <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>. Acesso em 18 jun. 2023.

[25] HANU, Lauro, THEWLIS, James e HACO, Sasha. *How AI is learning to identify toxic online content.* Scientific American, Estados Unidos, 8 de fevereiro de 2021. Disponível em: <https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/>. Acesso em 18 jun. de 2023.

[26] PINTO, Felipe Chiarello de Souza; GARCIA, Lara Rocha; ROSA, Alexandre Morais da. *Direito para humanos robotizados ou direito dos robôs humanizados*? Novos Estudos Jurídicos, v. 28, 2023, p. 542. Disponível em: <https://periodicos.univali.br/index.php/nej/article/view/19985>. Acesso em 24 fev. 2024.

[27] CHIARELLO, Felipe Chiarello; GARCIA, Lara Rocha. *Desafios internacionais da aplicação da inteligência artificial no Direito*. Justiça do Direito (UPF), v. 35, 2021, p. 15. Disponível em :<https://seer.upf.br/index.php/rjd/article/view/13040/114116181>. Acesso em 26 fev. 2024.

Uma das mais notáveis dificuldades reside na competência da IA em interpretar e contextualizar corretamente a linguagem. Os algoritmos de IA podem, por exemplo, confundir comentários prejudiciais com comentários que, embora contenham termos relacionados a gênero, orientação sexual, religião ou deficiência, sejam inofensivos.[28] Essa limitação é ecoada em pesquisas acadêmicas, que destacam os crescentes desafios de tornar esses algoritmos tanto previsíveis quanto explicáveis:

> Olhando para o futuro mais imediato, confrontamos duas dificuldades especialmente salientes à medida que os algoritmos de aprendizado se tornam mais sofisticados. São os problemas de "previsibilidade" e "explicabilidade". A previsibilidade de um algoritmo é uma medida de quão difícil é prever suas saídas, enquanto sua explicabilidade é uma medida de quão difícil é explicar suas saídas. Esses problemas são familiares para a comunidade de robótica, que há muito tempo busca lidar com a preocupação de que robôs possam interpretar comandos de forma muito literal (por exemplo, instruído para escurecer uma sala, o robô destrói as lâmpadas). Algoritmos de aprendizado abstrato se deparam diretamente com essa dificuldade. Mesmo que possamos descrever completamente o que os faz funcionar, os mecanismos reais pelos quais eles implementam suas soluções provavelmente permanecerão opacos: difíceis de prever e às vezes difíceis de explicar. E à medida que se tornam mais complexos e mais autônomos, essa dificuldade aumentará. (...)
>
> Os seres humanos são outro exemplo de um sistema frequentemente imprevisível e inexplicável. Utilizamos regras legais, incentivos, direitos e prerrogativas para mudar o comportamento humano. No entanto, às vezes é difícil saber antecipadamente se uma determinada intervenção social será eficaz e, mesmo que seja, se produzirá consequências não intencionadas. Mas uma diferença importante entre algoritmos de aprendizado de máquina e seres humanos é que os humanos têm uma vantagem inata ao tentar prever e explicar o comportamento humano. Ou seja, evoluímos para nos entendermos. Os humanos são criaturas sociais cujos cérebros evoluíram a capacidade de desenvolver teorias da mente sobre outros cérebros humanos. Não há uma vantagem natural semelhante para intuir como os algoritmos se comportarão.
>
> Determinar que um algoritmo é suficientemente previsível e explicável para ser "seguro" é difícil, tanto do ponto de vista técnico quanto do ponto de vista de políticas públicas. Se um algoritmo é insuficientemente previsível, ele pode ser mais perigoso do que sabemos. Se um algoritmo é insuficientemente explicável, pode ser difícil saber como corrigir suas saídas problemáticas. (grifos nossos) [29]

Estudos atuais comprovaram que é legítima a preocupação com a utilização majoritária da tecnologia e, sobretudo, da inteligência artificial para fins de desenvolvimento de produtos e serviços disponibilizados ao público. Levantamentos recentes, dentre os quais se destaca o realizado por pesquisadores do Instituto de Tecnologia da Georgia, nos Estados Unidos, reforçam a existência de desafios relacionados à aplicação prática da inteligência artificial. O objeto do estudo foi a capacidade dos sistemas de detecção de veículos autônomos[30] em identificar pedestres de

[28] HANU, Lauro, THEWLIS, James e HACO, Sasha. *How AI is learning to identify toxic online content Op. Cit.*

[29] No original: "Looking to the more immediate future, we confront two especially salient difficulties as learning algorithms become more sophisticated. They are the problems of 'predictability' and 'explainability.' An algorithm's predictability is a measure of how difficult its outputs are to predict, while its explainability is a measure of how difficult its outputs are to explain. Those problems are familiar to the robotics community, which has long sought to grapple with the concern that robots might misinterpret commands by taking them too literally (e.g., instructed to darken a room, the robot destroys the lightbulbs). Abstract learning algorithms run headlong into that difficulty. Even if we can fully describe what makes them work, the actual mechanisms by which they implement their solutions are likely to remain opaque: difficult to predict and sometimes difficult to explain.

And as they become more complex and more autonomous, that difficulty will increase. (…) Humans are another example of an often unpredictable and inexplicable system. We use legal rules, incentives, entitlements, and rights to change human behavior. Nevertheless, it is sometimes difficult to know in advance whether a given social intervention will be effective and, even if it is effective, whether it will produce unintended consequences. 10 7 But an important difference between machine-learning algorithms and humans is that humans have a built-in advantage when trying to predict and explain human behavior. 108 Namely, we evolved to understand each other. 109 Humans are social creatures whose brains have evolved the capacity to develop theories of mind about other human brains. There is no similar natural edge to intuiting how algorithms will behave. Determining that an algorithm is sufficiently predictable and explainable to be "safe" is difficult, both from a technical perspective and a public policy perspective. If an algorithm is insufficiently predictable, it could be more dangerous than we know. If an algorithm is insufficiently explainable, it might be difficult to know how to correct its problematic outputs. Indeed, it may be extremely difficult even to know what kinds of outputs are 'errors'". *In:* TUTT, Andrew. *An FDA for Algorithms*. Administrative Law Review, v. 69, n. 1, 2017, p. 101-103. Acesso em 31 ago. 2023. Disponível em <http://www.administrativelawreview.org/wp-content/uploads/2019/09/69-1-Andrew-Tutt.pdf>.

[30] Esses veículos, equipados com tecnologia que permite a condução sem a intervenção humana direta, dependem de uma série de sensores, câmeras e algoritmos para transitar.

diferentes tons de pele. Utilizando a escala de *Fitzpatrick*[31], um sistema de classificação de tons de pele humana, os pesquisadores compararam a acuracidade da captação de pessoas com pele clara em relação às pessoas com pele escura.[32] Os resultados mostraram uma tendência de menor precisão na detecção de indivíduos com tons de pele mais escuros,[33] e essa variação não estava relacionada a fatores como iluminação ou obstruções visuais.[34]

Em outro estudo similar, foi apurado que, além das falhas na detecção de pessoas com pele escura, os veículos autônomos também não identificam crianças com a acurácia desejada. A pesquisa analisou oito sistemas de detecção de pedestres baseados em inteligência artificial, comuns em estudos sobre veículos autônomos. Ao avaliar mais de 8.000 imagens, constatou-se que a eficácia na (i) detecção de adultos superava em quase 20% a de crianças, enquanto a (ii) identificação de indivíduos com pele clara era 7,5% maior do que pessoas com pele escura.[35]

E esses não são os únicos exemplos de que a inteligência artificial pode apresentar falhas graves. No Estados Unidos, um *software* de avaliação de risco denominado de *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) é utilizado pelo sistema judiciário para examinar a probabilidade de um detento reincidir no cometimento de algum crime. Em tese, os resultados da avaliação de risco indicariam aos magistrados o índice de periculosidade do réu à comunidade. No entanto, a organização de jornalismo investigativo *ProPublica* avaliou as pontuações de risco atribuídas a mais de 10.000 réus criminais no condado de Broward, na Florida, no período entre 2013 e 2014 e concluiu que o *software* é tendencioso contra os negros. Apurou-se que 45% (quarenta e cinco por cento) das pessoas negras não reincidentes nesses dois anos foram equivocadamente classificadas como de alto risco pelo *software*, enquanto somente 23% (vinte e três por cento) das pessoas brancas que não reincidiram foram categorizadas da mesma forma.[36] Em contrapartida, os réus brancos "que reincidiram nos dois anos seguintes foram erroneamente rotulados de baixo risco quase duas vezes mais do que os reincidentes negros (48% vs. 28%)".[37]

Diante dessas limitações, ao menos, por ora, é improvável que a inteligência artificial, por si só, seja capaz de remover adequadamente postagens com teor nocivo na internet. Para ilustrar, imagine-se a hipótese de um portal midiático que apenas divulgou uma notícia sobre terceiros que proferiram um discurso antidemocrático. Nesse contexto, apesar do caráter meramente informativo da publicação, o uso exclusivo da inteligência artificial, sem o auxílio fundamental de um moderador de conteúdo humano, poderia resultar em remoções indevidas de conteúdo.

Portanto, as redes sociais não devem relegar a responsabilidade da moderação de conteúdo exclusivamente à inteligência artificial. Em vez disso, uma estratégia mais ponderada e com maior presença humana pode se revelar mais eficaz. Isso pode envolver a integração da inteligência artificial com equipes de moderação compostas por pessoas, além de proporcionar aos usuários a possibilidade de interagir com profissionais reais, em vez de apenas com sistemas automatizados. Uma abordagem deste tipo, mais personalizada e contextualizada, poderia ajudar a lidar com certas limitações da inteligência artificial, enquanto ainda se aproveita de suas vantagens no controle do conteúdo.

Ao contrário do que se imagina, na hipótese de bem estruturada determinada política pública para gestão e controle da atividade dessas plataformas digitais e sua interface com seus usuários, a fim de que apenas sejam

[31] A Escala de Fitzpatrick, desenvolvida pelo médico norte-americano Thomas B. Fitzpatrick, é um método de categorização internacional utilizado na área dermatológica para classificar distintos tipos de pele humana com base em sua resposta à exposição solar (SBD - Sociedade Brasileira de Dermatologia. *Classificação dos fototipos de pele*, 2021. Acesso em 15 jan. 2024. Disponível em: <https://www.sbd.org.br/cuidados/classificacao-dos-fototipos-de-pele/>).

[32] WILSON, Benjamin. HOFFMAN, Judy. MORGENSTERN, Jamie. *Predictive Inequity in Object Detection*, arXiv:1902.11097v1 [cs.CV], 2019. Acesso em 17 set. 2023. Disponível em <https://arxiv.org/pdf/1902.11097.pdf>, p. 3-4.

[33] *Ibidem*, p. 6.

[34] *Ibidem*, p. 8.

[35] LI, Xinyue; CHEN, Zhenpeng; ZHANG, Jie M; SARRO; Federica; ZHANG, Ying; LIU, Xuanzhe. *Dark-Skin Individuals Are at More Risk on the Street: Unmasking Fairness Issues of Autonomous Driving Systems*, arXiv:2308.02935v1 [cs.CY]. Reino Unido, 2023, p. 6. Disponível em <https://arxiv.org/pdf/2308.02935.pdf>. Acesso em 17 set. 2023.

[36] ANGWIN. Julia, LARSON, Jeff. MATTU, Surya. KIRCHNER, Lauren. *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublic, 2016. Disponível em <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. Acesso em 24 set. 2023.

[37] No original: "Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent)".

respeitadas aquelas diretrizes previamente estabelecidas pelos próprios provedores de aplicação, parece ser palatável a ideia de proteção dos valores democráticos mediante a dispensa de uma atuação massiva e contundente do Poder Judiciário e do Congresso Nacional, bastaria para isso que houvesse a estruturação, pelo Executivo, de mecanismos de controle extrajudicial do relacionamento entre usuários e provedores de aplicação. Para tanto, é preciso dar alguns passos para trás, a fim de, na contramão daquilo que pretendem as redes sociais, humanizar o tratamento das denúncias, afastando essa pretensa modernização mediante a adoção de inteligência artificial e sistemas autônomos para tratamento das reclamações. Evidentemente, essa estratégia envolve questões polêmicas que devem ser avaliadas pelo Poder Público, como a própria exigência de prévia instalação de sede no Brasil desses provedores, em especial para que seja possível o atendimento aos usuários no idioma oficial.

Definitivamente, a Sociedade Civil, os Agentes Privados e o Poder Público não estão alheios a essa discussão, o que se atesta pela urgência (RICD,[38] art. 154) atribuída ao regime de tramitação do Projeto de Lei n.º 2.630/2020, popularmente conhecido como PL das *Fake News*, mas formalmente denominado como projeto da "Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet".

O combate às contas fraudulentas ou à disseminação de informações falsas parece ter norteado a preocupação dos parlamentares (art. 6º, incs. I, II, III, art. 7º, parágrafo único, art. 14, incs. I e II, art. 15, incs. I, II, III, IV e V, art. 16, art. 17, entre outros). Dentre as mais de 150 emendas parlamentares no Senado Federal e as 85 já existentes na Câmara dos Deputados em relação a esse projeto legislativo, uma delas merece especial destaque neste trabalho, trata-se da EMP[39] n.º 1 em Plenário da Câmara dos Deputados, datada de 5 de abril de 2022. Com ela, o Dep. Mário Heringer, do PDT-MG, demonstrou precisamente a mesma prudência aqui defendida, no sentido de que o atendimento a ser prestado pelas redes sociais deve ser humanizado. Malgrado aqui se entenda pela mesma cautela, mas por razões diversas daquelas apontadas pelo parlamentar em sua justificativa de motivos,[40] a necessidade de humanizar o tratamento das denúncias formalizadas é um dado e teve sua necessária atenção do legislador, seja para fins de o usuário reclamar sobre bloqueio indevido ou então para que este mesmo usuário persiga determinado bloqueio de conteúdo indevido.

Essa discussão sobre a humanização do tratamento após formalizadas queixas ou denúncias não ofusca uma polêmica criação legislativa que esteve em voga durante a tramitação do PL das *Fake News*. Trata-se da criação de uma espécie de órgão fiscalizador, o Conselho de Transparência e Responsabilidade na Internet.

Essa figura esteve presente até a divulgação do parecer preliminar pelo relator do PL na Câmara dos Deputados, em abril de 2023. Apesar dessa iniciativa ter sido extirpada do projeto que ainda segue em análise na Câmara dos Deputados, este órgão teria sido idealizado para elaborar espécie de código de conduta para as redes sociais, para

[38] Regimento Interno da Câmara dos Deputados.

[39] Emenda de Plenário.

[40] "O PL 2.630/2020 avança em variadas temáticas relevantes para se fortalecer o enfrentamento da desinformação na internet. A proposição estabelece regras equilibradas para se evitar que o combate ao espalhamento de boatos não se converta em censura. No espírito da proposição, esta emenda busca estabelecer o adequado atendimento aos usuários que tenham seu conteúdo bloqueado por suspeita de violação dos termos de uso dos provedores de redes sociais, ferramentas de busca e de serviços de mensageria instantânea. O Art. 15 da proposição já estabelece regras para se garantir a razoabilidade da aplicação desses termos de uso e a adequada notificação do usuário caso um conteúdo por ele publicado supostamente viole tais termos. É razoável que, dada a enorme quantidade de conteúdo publicado diariamente, os provedores utilizem sistemas automatizados para detectar publicações suspeitas de propagar desinformação ou afrontar determinado direito subjetivo. Porém, tais sistemas informatizados eventualmente cometem erros de avaliação, e terminam bloqueando conteúdos que não violam termos de uso, nem propagam desinformação, nem afrontam direitos. Esse entendimento sobre a adequação do conteúdo injustamente bloqueado deve ser emitido após avaliação por seres humanos, pois trata-se de julgamento de adequação moral impossível de se ensinar para uma máquina. É fato que alguns provedores de redes sociais não apenas realizam a detecção de conteúdo suspeito por meio de sistemas automatizados, mas também realizam o atendimento aos usuários injustamente bloqueados exclusivamente por sistemas automatizados. Por esta razão, a presente emenda busca obrigar os provedores a fornecerem atendimento por seres humanos, vedando o atendimento por sistemas automatizados, aos usuários que desejem reclamar de um bloqueio injusto. O atendimento realizado exclusivamente por sistemas automatizados pode se converter na inviabilização da solução administrativa de lides tocantes ao bloqueio de conteúdos e, indiretamente, se converter em autorização para o exercício do poder de censura pelos provedores. Além disso, a emenda prevê que os provedores devem responder até vinte e quatro horas os pedidos referentes a publicações com impulsionamento pago. Com isto, busca-se evitar que publicações ofensivas impulsionadas produzam enorme dano à pessoa ofendida e, ao mesmo tempo, evita-se que o provedor restrinja injustamente o potencial de uma publicação paga com conteúdo justo e relevante. Logo, para que se preserve o espírito de equilíbrio do PL 2.630/2020, rogo ao relator e aos pares a aprovação desta emenda" (*justificativa de motivos*, EMP n.º 1, de 5 de abril de 2022, ao PL 2630/2020, Câmara dos Deputados).

avaliar as políticas de uso adotadas pelos provedores de aplicação, além dos procedimentos de moderação adotados pelos provedores de redes sociais, entre várias outras importantes atribuições.

Dada a delicadeza do tema e das atividades que seriam desempenhadas pela entidade fiscalizadora, a ideia não vingou até agora (durante a redação destas breves ideias). Contudo, isso não impede que comentários sejam tecidos sobre essa proposição legislativa que, em algum momento, esteve em forma de ato legislativo, ainda que apenas *em potência*. Até porque, como o projeto de lei ainda não foi aprovado, não se descarta a possibilidade de um novo debate sobre esse tema.

O órgão deliberativo originalmente idealizado seria composto de 21 (vinte e um) conselheiros, com mandato de 2 (dois) anos, sendo admitida 1 (uma) recondução. Se por um lado a criação desse órgão parece assustar por seu possível perfil autoritário, a composição originalmente desenhada pelo PL reflete uma intenção claramente democrática, pois envolveria representantes do Senado Federal, da Câmara dos Deputados, do Conselho Nacional de Justiça, do Conselho Nacional do Ministério Público, do Comitê Gestor da Internet no Brasil, da própria sociedade civil, da academia e comunidade técnica, dos provedores de acesso, aplicações e conteúdo da internet, dos setores de comunicação social, telecomunicações, do Conselho Nacional dos Chefes de Polícia Civil, do Departamento de Polícia Federal, da Agência Nacional de Telecomunicações e, por fim, do Conselho Nacional de Autorregulamentação Publicitária.

Talvez pela incerteza que circulava acerca da eficácia da Lei Geral de Proteção de Dados, a proposição desse órgão tal qual lançada anteriormente não contemplava a participação da Agência Nacional de Proteção de Dados – LGPD, que foi instituída apenas em 2018, com a promulgação da Lei Geral de Proteção de Dados Pessoais. No entanto, em havendo a criação de um órgão com finalidade de assegurar transparência e responsabilidade no uso da internet, parece certo dizer que qualquer deliberação a respeito do que pode ou não estar presente no mundo virtual também precise passar pela avaliação de especialistas dessa nova agência reguladora responsável pela proteção de dados. Desse modo, na hipótese dessa discussão voltar à tona no Parlamento, é recomendável que nesse Conselho seja assegurada a representação também de especialistas da Agência Nacional de Proteção de Dados.

Havendo ou não um órgão próprio para deliberar sobre o conteúdo *online*, isso não desonerará as plataformas digitais de aperfeiçoarem seus mecanismos de controle, sobretudo para que seja dada maior efetividade ao índice de resolução de queixas e denúncias formalizadas por usuários.

CONSIDERAÇÕES FINAIS

Como se demonstrou nas linhas acima, está longe de ter um fim o debate acerca dos limites da liberdade de expressão imerso no mundo virtual. Ao contrário, a impressionante velocidade com que avançam as descobertas tecnológicas é exatamente a mesma que impulsiona a tensão entre grandezas tão relevantes para o cenário global: liberdade de expressão e preservação democrática. Se de um lado a vivência histórica pode conduzir a um modelo regulador mais ativo e protetivo de valores democráticos, é verdade que um outro contexto é capaz de produzir um modelo menos regulado, em que se preserve a competição livre de ideais, numa busca saudável (ou nem tanto assim) pela verdade.

É verdade que a definição do grau de intervenção do Estado naquilo que se veicula no mundo virtual vai depender muito desse modelo que virá a ser adotado. No entanto, uma das conclusões que se extrai destas breves reflexões é que o papel das plataformas digitais – os *provedores de aplicação*, como denomina a legislação brasileira – pode ser mais ativo e não deve ignorar a relevância do canal informativo que estas *big techs* têm sob seu poder. Nesse sentido, parece-nos que não basta que as diretrizes e os termos de consentimento sejam extensos, ou mesmo que a fiscalização do conteúdo postado seja terceirizada unicamente à inteligência artificial, que deve ser uma ferramenta valiosa no combate à desinformação, ao discurso de ódio e às *fake news*, mas não a única alternativa.

Desse modo, parece-nos relevante que a implementação de políticas públicas por parte do Estado passe pela reavaliação acerca da exigência de procedimentos internos mais transparentes na resolução de queixas e denúncias por parte das plataformas digitais, o que envolve logicamente a interveniência humana, que contará com o auxílio da inteligência artificial na condição de mero coadjuvante (e não protagonista). E é justamente este o grande desafio para a proteção da Democracia no século XXI: reconhecer a necessidade de dar alguns passos atrás no avanço tecnológico, para se perceber que a construção de um ambiente mais saudável no mundo virtual talvez demande a humanização do tratamento de queixas e denúncias formalizadas às plataformas digitais, embora não se desconheça o quão custoso possa vir a se tornar essa tarefa às *big techs*.

REFERÊCIAS BIBLIOGRÁFICAS

ANGWIN. Julia, LARSON, Jeff. MATTU, Surya. KIRCHNER, Lauren. *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublic, 2016. Disponível em <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. Acesso em 24 set. 2023.

BERLIN, Isaiah. *Quatro ensaios sobre a liberdade*. Trad. Wumberto Hudson Ferreira. Brasília, DF: Ed. UnB, 1981.

BINENBOJM, Gustavo; PEREIRA NETO, Caio Mário da Silva. Prefácio. *In*: FISS, Owen M. *A Ironia da Liberdade de Expressão: Estado, Regulação e Diversidade na Esfera Pública*. Rio de Janeiro: Renovar, 2005.FILHO, Adalberto Camargo Aranha; PINTO, Felipe Chiarello de Souza; RAMOS, Tais. *A Liberdade de Expressão Política e o Discurso de Ódio (HateSpeech) em Período Eleitoral: estudos de casos da Corte Interamericana de Direitos Humanos*. *In*: MARQUES, Claudia Lima; MARTINI, Sandra Regina; FINCO, Matteo (org). *Diálogos entre direitos humanos, direito do consumidor, compliance e combate à corrupção*. 1. ed. São Paulo: YKEditora, 2021, p. 155.

CHIARELLO, Felipe; GARCIA, Lara Rocha. *Desafios internacionais da aplicação da inteligência artificial no Direito*. Justiça do Direito (UPF), v. 35, 2021, p. 15. Disponível em :<https://seer.upf.br/index.php/rjd/article/view/13040/114116181>. Acesso em 26 fev. 2024.

FREITAS, Riva Sobrado de; CASTRO, Matheus Felipe de. *Liberdade de Expressão e Discurso de Ódio: um exame sobre as possíveis limitações à liberdade de expressão*. Rev. Seqüência (Florianópolis), 66, p. 327-355, jul.2013. Disponível em <https://doi.org/10.5007/2177-7055.2013v34n66p327>. Acesso em 5 jul. 2023.

HANU, Lauro, THEWLIS, James e HACO, Sasha. *How AI is learning to identify toxic online content.* Scientific American, Estados Unidos, 8 de fevereiro de 2021. Disponível em: <https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/>. Acesso em 18 jun. de 2023.

LI, Xinyue; CHEN, Zhenpeng; ZHANG, Jie M; SARRO; Federica; ZHANG, Ying; LIU, Xuanzhe. *Dark-Skin Individuals Are at More Risk on the Street*: Unmasking Fairness Issues of Autonomous Driving Systems, arXiv:2308.02935v1 [cs. CY]. Reino Unido, 2023. Disponível em <https://arxiv.org/pdf/2308.02935.pdf>. Acesso em 17 set. 2023.

MACEDO JUNIOR, Ronaldo Porto Macedo. *Freedom of Expression: what lessons should we learn from US experience?* São Paulo: Revista Direito GV, v. 13, n. 1, págs. 274-302, jan-abr/2017. Disponível em <https://www.scielo.br/j/rdgv/a/tRnqx97GRkqny4L77JFGBTx/?format=pdf&lang=en>. Acesso em 6 jun. 2023.

MILL, John Stuart. *On Liberty* (1859). Kitchener: Batoche Books, 2001.

NEWTON, Casey. *Bodies in seats.* he Verge: Vox Media, Estados Unidos, 19 de junho de 2019. Disponível em: <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>. Acesso em 18 jun. 2023.

PINTO, Felipe Chiarello de Souza; GARCIA, Lara Rocha; ROSA, Alexandre Morais da. *Direito para humanos robotizados ou direito dos robôs humanizados?* Novos Estudos Jurídicos, v. 28, 2023, p. 542. Disponível em: <https://periodicos.univali.br/index.php/nej/article/view/19985>. Acesso em 24 fev. 2024.

SARMENTO, Daniel. *A liberdade de expressão e o problema do hate speech*. Revista de Direito do Estado, Rio de Janeiro, v. 1, n. 4, p. 53-105, out.-dez. 2006. Disponível em: <www.dsarmento.adv.br/content/3-publicacoes/18-a-liberdade-de-expressao-e-oproblema-do-hate-speech/a-liberdade-de-expressao-e-o-problema-do-hate-speech-daniel-sarmento.pdf>. Acesso em 17 ago. 2023.

SOUZA, Carlos Affonso; LEMOS, Ronaldo. *Marco civil da internet: construção e aplicação*. Juiz de Fora: Editar Editora Associada Ltda., 2016.

STONE, Geoffrey R. *Free Speech in the Twenty-First Century: Ten Lessons from the Twentieth Century Lead Article*. Rev. Pepperdine Law Review, v. 36, Issue 2, mai-2009, págs. 273-300. Disponível em <https://digitalcommons.pepperdine.edu/cgi/viewcontent.cgi?article=1072&context=plr >. Acesso em 6 jun. 2023.

TUTT, Andrew. *An FDA for Algorithms*. Administrative Law Review, v. 69, n. 1, 2017, p. 101-103. Disponível em <http://www.administrativelawreview.org/wp-content/uploads/2019/09/69-1-Andrew-Tutt.pdf>. Acesso em 31 ago. 2023.

WALDRON, Jeremy. *Dignity and defamation: the visibility of hate*. Harvard Law Review, v. 123, n. 1.596, p. 1.597-1.657, 2010. LEAL DA SILVA, Rosane *et al*. *Discursos de ódio em redes sociais*: jurisprudência brasileira. Revista FGV Law, São Paulo, v. 7, n. 2, p. 445-468, jul.-dez.2011.

WILSON, Benjamin. HOFFMAN, Judy. MORGENSTERN, Jamie. *Predictive Inequity in Object Detection*, arXiv:1902.11097v1 [cs.CV], 2019. Disponível em <https://arxiv.org/pdf/1902.11097.pdf>. Acesso em 17 set. 2023.

WOISCHNIK, Jan. (*ed.*); MARTINS, Leonardo (*org.*). *Cinquenta anos de jurisprudência do Tribunal Constitucional Federal Alemão*. Trad. Beatriz Hennig, Leonardo Martins, Mariana Bigelli de Carvalho, Tereza Maria de Castro e Vivianne Geraldes Ferreira. Konrad-Adenauer-Stiftung E.V., 2005.