v. 19 n. 32 São Paulo Dez. 2022

sur

revista internacional
de direitos humanos

edição **32**

ARTIGO

INTELIGÊNCIA ARTIFICIAL E MODERAÇÃO DO DISCURSO DE ÓDIO *ON-LINE*

Natalie Alkiviadou

• *Uma combinação arriscada?* •

RESUMO

As plataformas de mídias sociais têm cada vez mais feito uso da inteligência artificial para combater discursos de ódio on-line. A imensa quantidade de conteúdos, a velocidade em que eles são desenvolvidos e a crescente pressão estatal sobre as empresas para rapidamente remover discursos de ódio de suas plataformas levaram a uma situação complicada. A presente análise argumenta que os mecanismos automatizados, por poderem apresentar conjuntos de dados enviesados e, desse modo, serem incapazes de identificar as nuances da linguagem, não deveriam ser deixados sem precaução com o discurso de ódio, uma vez que isso pode provocar violações da liberdade de expressão e do direito à não discriminação.

PALAVRAS-CHAVE

Liberdade de expressão | Discurso de ódio | Inteligência artificial | Plataformas de mídias sociais

1 • Introdução

As plataformas de mídias sociais (SMP, no original em inglês, ou PMS, na tradução ao português) constituem uma das principais fontes de comunicação e de informação. Elas facilitam a comunicação sem fronteiras, permitem diversas formas de expressão – como a política, a ideológica, a cultural e a artística –, dão voz a grupos tradicionalmente silenciados, oferecem uma alternativa à mídia mainstream (que pode estar sujeita à censura do Estado), possibilitam a difusão de notícias diárias e promovem a conscientização sobre violações de direitos humanos. Contudo, conforme observado por Mchangama *et al.*,[1] o uso massivo das PMSs dá uma nova visibilidade a fenômenos como o ódio e o abuso. A utilização das PMSs também está diretamente vinculada a eventos tenebrosos como o genocídio em Mianmar. Ao reconhecer os perigos de um discurso violento como um risco iminente de violência, argumenta-se que se deve ter precaução ao abraçar a retórica comum de que o discurso de ódio prevalece através das mídias sociais, uma vez que o trabalho empírico tem demonstrado o contrário. Por exemplo, Siegel *et al.* conduziram um estudo para avaliar se a campanha eleitoral de [Donald] Trump em 2016 (e no período de seis meses subsequentes) gerou um aumento do discurso de ódio no Twitter.[2] Com base na análise de uma amostra de 1,2 bilhão de tweets, constatou-se que entre 0,001 e 0,003% dos tweets continha discurso de ódio em qualquer que fosse o dia – "uma fração diminuta tanto do conteúdo com teor político quanto dos conteúdos em geral produzidos por estadunidenses que usam o Twitter".

Ainda assim, a pressão estatal pela regulamentação do discurso de ódio nas plataformas está aumentando, o que, conforme discorrido neste artigo, levou à diluição do direito à liberdade de expressão e contribuiu diretamente para o silenciamento de grupos minoritários. A forma como essa nova realidade está sendo encarada por Estados e instituições, como a União Europeia, é preocupante. Por exemplo, em 2017, a Alemanha aprovou a Network Enforcement Act (Lei de Fiscalização de Redes – NetzDG, no original em inglês), que busca combater discursos on-line ilegais, como insultos, incitação e difamação religiosa. Por meio da força dessa lei, as plataformas de mídias sociais com mais de 2 milhões de pessoas usuárias ativas são obrigadas a remover conteúdos ilegais – incluindo discursos de ódio e ofensas religiosas – no prazo de 24 horas; caso contrário, arriscam-se a pagar multas pesadas de até 50 milhões de euros. Tal ação se transformou em um protótipo para a governança da internet em Estados autoritários. Em dois relatórios de Mchangama *et al.*, um de 2019 e outro de 2020, a organização Justitia registrou a aprovação de um modelo da lei NetzDG em mais de 20 países, muitos dos quais haviam sido classificados pela organização Freedom House como "não livres" ou "parcialmente livres".[3] Todos os países exigem que as plataformas *on-line* removam categorias vagas de conteúdo, o que inclui "informações falsas", "blasfêmia/insulto religioso" e "discurso de ódio". Mchangama e Alkiviadou observam com preocupação que "poucos desses países têm em vigor um Estado de direito básico e proteções à liberdade de expressão incorporados a exemplo do precedente alemão".[4] Um modelo similar tem sido seguido no âmbito da União Europeia (UE) por meio da Lei de Serviços Digitais (DSA, no original em inglês).[5]

Como medida de resposta às exigências regulatórias atualizadas, levando-se em consideração as pesadas multas, as plataformas encontram-se inclinadas a adotar a abordagem "melhor prevenir do que remediar" e a regular os conteúdos com rigidez. Entretanto, conforme apontado por Llanso,[6] a comunicação on-line nessas plataformas ocorre em uma escala massiva, tornando impossível que uma moderação humana possa analisar todos os conteúdos antes de eles se tornarem disponíveis. Ainda, a imensa quantidade de conteúdos *on-line* também torna o trabalho de análise, ainda que apenas dos conteúdos denunciados, uma tarefa árdua. Para atender à necessidade de esquivar-se das multas do governo e dos aspectos técnicos de escala e quantidade de conteúdos, as PMSs têm dependido cada vez mais da inteligência artificial (IA), na forma de mecanismos automatizados para abordar os problemas de conteúdo de forma proativa ou reativa, inclusive discursos de ódio. Em resumo, assim como destacado por Dias *et al.*,[7] a IA fornece às PMSs "ferramentas para policiar um fluxo enorme e crescente de informações – o que vem a ajudar a implementação de políticas de conteúdo". Embora a IA seja necessária em áreas que envolvem, por exemplo, o abuso infantil e a divulgação não consensual de atos de intimidade entre adultos, seu emprego para regular áreas nebulosas e mais controversas do discurso humano, como é o caso do discurso de ódio, é complexo. Em vista desses desenvolvimentos tecnológicos, este artigo concentra-se no uso da IA para a regulação do discurso de ódio nas PMSs, argumentando que mecanismos automatizados, por poderem apresentar conjuntos de dados enviesados e, desse modo, serem incapazes de identificar as nuances da linguagem, podem gerar violações da liberdade de expressão e do direito à não discriminação de grupos minoritários, dessa forma silenciando ainda mais grupos já marginalizados.

2 • Discurso de ódio: noções e semântica

Não existe uma definição universalmente aceita para discurso de ódio. A maioria dos Estados e instituições adotam um entendimento próprio do que o discurso de ódio implica,[8] sem defini-lo.[9] Um dos poucos documentos, ainda que não vinculante, que buscaram explicar o significado do termo é a Recomendação do Comitê de Ministros do Conselho Europeu sobre discurso de ódio.[10] Ela dispõe que o termo deve ser

> *compreendido de forma a abranger todas as formas de expressão que disseminem, incitem, promovam ou justifiquem o ódio racial, a xenofobia, o antissemitismo ou outras formas de ódio com base na intolerância, inclusive expressões intolerantes na forma de nacionalismo ou etnocentrismo agressivo, discriminação e hostilidade contra minorias, migrantes e pessoas de origem imigrante.*

O discurso de ódio também foi citado, porém não definido, pelo Tribunal Europeu de Direitos Humanos (TEDH). A título ilustrativo, o tribunal constatou que o discurso de ódio envolve "todas as formas de expressão que disseminem, incitem, promovam ou justifiquem o ódio com base na intolerância, inclusive a intolerância religiosa".[11] Adicionar o trecho sobre

a justificativa do ódio por si só demonstra o baixo limiar para um discurso ser considerado inaceitável. Ademais, em suas decisões, o TEDH proferiu que, para ser considerado um discurso de ódio, não é necessário que o discurso "incite indivíduos diretamente para que estes cometam atos de ódio",[12] pois ataques contra pessoas podem ser praticados por meio de "insulto, ridicularização ou crime contra a honra de grupos específicos da população"[13]e "o discurso empregado de maneira irresponsável não pode ser digno de proteção".[14] Nesse sentido, o TEDH delineou a correlação entre discurso de ódio e seus efeitos negativos sobre as vítimas, alegando que mesmo discursos não violentos e equivalentes a meros insultos têm o potencial de causar danos suficientes para justificar o cerceamento da liberdade de expressão.

Além disso, a Agência de Direitos Fundamentais da UE preparou duas formulações distintas para discurso de ódio, sendo que a primeira "se refere ao incitamento e incentivo ao ódio, discriminação ou hostilidade contra um indivíduo, ação essa motivada pelo preconceito contra essa pessoa em razão de uma dada característica".[15] Em seu relatório de 2009 sobre a homofobia, a agência afirmou que o termo discurso de ódio, conforme utilizado na referida seção do relatório, "inclui um espectro mais amplo de atos de fala, englobando discursos públicos desrespeitosos".[16] A parte especificamente problemática dessa definição é a referência abrangente a discursos públicos desrespeitosos, principalmente porque instituições como o TEDH estendem a liberdade de expressão a ideias que "chocam, ofendem ou perturbam".[17] Este é o posicionamento formal do tribunal, mesmo que, no que tange aos casos de discurso de ódio, conforme brevemente observado anteriormente, ele tenha rigorosamente adotado um limite muito baixo do que ele está disposto a aceitar como discurso admissível.

Direcionando agora o foco para as plataformas, embora esteja além do escopo deste artigo avaliar todas as diretrizes e normas que recaem sobre as PMSs, contemplaremos duas abordagens distintas: o Facebook e o Instagram, de um lado (ambos sob a propriedade da Meta Platforms Inc..), e o Reddit, de outro. Os primeiros[18] estabelecem um entendimento de discurso de ódio em três níveis: o primeiro é o do discurso violento e desumano; o segundo consiste em declarações de inferioridade, desprezo, rejeição e outras formas de "ofensa", como a repulsa; e o terceiro inclui declarações de teor segregacionista e excludente. A lista das características protegidas é vasta e apresenta aspectos como raça, etnia, afiliação religiosa, casta, orientação sexual e doença grave.[19] O Reddit[20] adota uma abordagem mais protetora do discurso, proibindo a incitação à violência e a promoção do ódio. As características sob proteção são raça, cor, religião e gravidez, entre outras. É válido ressaltar que todas as principais plataformas, incluindo as supracitadas e o Twitter,[21] o YouTube[22] e o TikTok,[23] incorporam os parâmetros de raça e religião na lista de características protegidas.

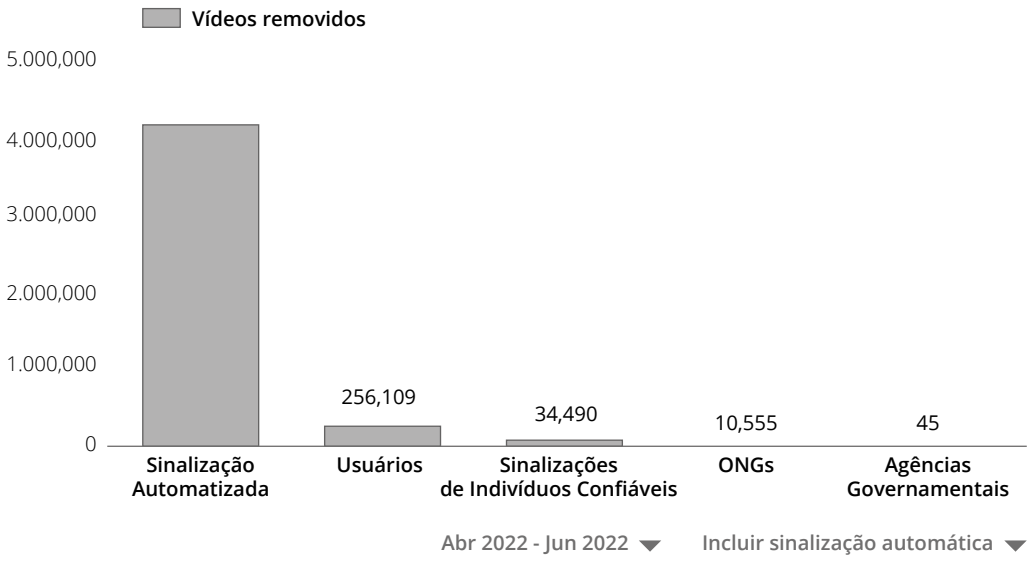
3 • Inteligência artificial

A aplicação da IA traduz-se em uma resposta à crescente pressão estatal sobre as plataformas de mídias sociais para que haja a retirada de discursos de ódio de modo rápido e eficiente. As PMSs também enfrentam a pressão de outros entes, como

anunciantes e pessoas usuárias da plataforma. Visando cumprir as normas (e evitar multas vultosas), as empresas fazem uso da IA, isoladamente ou em conjunto com uma moderação humana, para apagar supostos conteúdos de ódio. Conforme apontado por Dias, tais circunstâncias induziram as empresas a "agir proativamente para evitar a responsabilidade... em uma tentativa de proteger os modelos de negócio."[24]

Para dar um exemplo sobre o uso de IA pelas plataformas de mídias sociais, é possível comparar as taxas proativas de remoção de discurso de ódio do primeiro trimestre de 2018 (de 38%) e do segundo trimestre de 2022 (de 95,6%). Conforme observado em uma publicação do site Transparency Center, "nossa tecnologia detecta e remove proativamente a ampla maioria dos conteúdos violadores antes mesmo que alguém faça uma denúncia."[25]

Em seu último relatório sobre cumprimento das diretrizes[26] (segundo trimestre de 2022), o YouTube inseriu a ilustração a seguir como forma de demonstrar a porcentagem de sinalizações humanas e automatizadas com relação ao conteúdo removível (não apenas discurso de ódio):



Dias *et al.* expõem que os algoritmos desenvolvidos para conquistar essa automatização costumam ser personalizados por tipo de conteúdo, como imagens, vídeos, áudio e texto.[27] Segundo os achados de Duarte e Llanso,[28] as tecnologias atuais detectam textos nocivos por meio do processamento de linguagem natural e da análise de sentimentos e, ainda que tenham evoluído significativamente, sua precisão está entre 70 e 80%. Eles relatam que a IA conta com uma "capacidade limitada para analisar as nuances de significado presentes na comunicação humana ou para identificar a intenção ou motivação do locutor". Logo, tais tecnologias "ainda falham ao tentar compreender o contexto, imputando, portanto, riscos a quem utiliza as plataformas em termos de liberdade de expressão, acesso à informação e

igualdade". Além disso, Dias *et al*. discorrem que a transição da política ao código pode ocasionar mudanças de significado, uma vez que o código de máquina é mais limitado do que a sua contraparte humana.[29] Dado o poder que as PMSs detêm sobre o mercado atual de expressão e informação e sobre a necessidade e tendência crescentes de usar IA para lidar com pressões externas de remoção de conteúdo e com o volume de material, Cowls et al. afirmam que há uma demanda urgente por garantir que a moderação de conteúdo transcorra de modo a proteger os direitos humanos e o discurso público.[30]

A luz do exposto acima, e tendo em vista a área controversa do discurso de ódio, este estudo examinará os riscos aos direitos humanos decorrentes ou que possam porventura surgir a partir do atual *status quo* – ou seja, o aumento do uso da IA por empresas privadas com fins lucrativos –, concentrando a atenção na liberdade de expressão e na não discriminação.

4 • IA, discurso de ódio e desafios à liberdade de expressão

O Artigo 19 da Declaração Universal dos Direitos Humanos (DUDH) estabelece que "[t]odo ser humano tem direito à liberdade de opinião e expressão; este direito inclui a liberdade de, sem interferência, ter opiniões e de procurar, receber e transmitir informações e ideias por quaisquer meios e independentemente de fronteiras".

Esse direito à liberdade também se encontra protegido por outros documentos relevantes, como o Artigo 19 do Pacto Internacional sobre Direitos Civis e Políticos (PIDCP) e o Artigo 10 da Convenção Europeia de Direitos Humanos. Ambos os artigos contêm restrições à liberdade de expressão, ao passo que o Artigo 20 da PIDCP dispõe que:

> *1 - Toda a propaganda em favor da guerra deve ser interditada pela lei.*
> *2 - Todo o apelo ao ódio nacional, racial ou religioso que constitua uma incitação à discriminação, à hostilidade ou à violência deve ser interditado pela lei.*

Conforme Dias destaca,[31] confiar na IA, mesmo sem a supervisão humana, é uma necessidade no que concerne aos conteúdos que nunca poderiam ser ética ou legalmente justificáveis, como o abuso infantil. Entretanto, a questão torna-se complicada quando se trata de partes contestáveis do discurso, como o discurso de ódio, pois não há um posicionamento universal ético ou legal sobre o seu significado ou as ocasiões em que deve ser removido (se é que deveria). No âmbito desse tipo de discurso, Llanso salienta que o emprego de IA levanta "dúvidas relevantes sobre a influência da IA sobre o ambiente da informação e sobre, em última instância, os nossos direitos à liberdade de expressão e ao acesso à informação".[32] Conforme apontado por Llanso *et al.*,[33] o discurso apresenta "desafios distintos para a liberdade de expressão e o acesso à informação *on-line*". Um relatório do Conselho Europeu destaca que a aplicação da

IA para a regulação do discurso de ódio afeta diretamente a liberdade de expressão, o que gera preocupações sobre o Estado de Direito e, principalmente, sobre as noções de legalidade, legitimidade e proporcionalidade.[34] O Conselho Europeu constatou que o uso avançado de IA para moderar conteúdos pode resultar em bloqueios excessivos e, consequentemente, colocar em risco a liberdade de expressão.[35] Gorwa *et al.* afirmam que o aumento do uso de IA ameaça exacerbar a opacidade já existente na moderação de conteúdo, além de atrapalhar ainda mais a questão da justiça *on-line* e "turvar novamente a natureza fundamentalmente política das decisões de discurso feitas em escala".[36] Ainda, não obstante as especificações técnicas de um mecanismo específico, a identificação (e remoção) proativa do discurso de ódio constitui uma restrição prévia do discurso, com todos os problemas legais que isso implica. Especificamente, Llanso *et al.* defendem que há uma "forte premissa contra a validade da censura prévia na legislação internacional que versa sobre os direitos humanos".[37] O Ex-Relator Especial da ONU para a Liberdade de Opinião e Expressão, David Kaye, manifestou a sua preocupação quanto ao uso de ferramentas automáticas em termos de potenciais bloqueios excessivos e afirmou que as reivindicações para expandir a filtragem de *uploads* para áreas de conteúdo relacionadas a terrorismo e outros assuntos "ameaçam instituir um regime abrangente e desproporcional de censura prévia à publicação".[38]

5 • IA e desafios à não discriminação

Dias relata que o emprego de IA pode resultar na aplicação tendenciosa dos termos de serviço das empresas.[39] Isso provavelmente se deve à falta de dados e/ou a um conjunto de dados de treinamento enviesados, levando a um possível silenciamento de integrantes de grupos minoritários[40] e podendo causar violações da liberdade de expressão e do direito à não discriminação. Em seu relatório "Mixed Messages: The Limits of Automated Social Content Analysis" [Mensagens mistas: Os Limites da Análise Automatizada de Conteúdo Social, em tradução livre], o Center for Democracy and Technology revelou que mecanismos automatizados podem afetar o discurso de grupos marginalizados de forma desproporcional.[41] Embora tecnologias como o processamento de linguagem natural e a análise de sentimentos tenham sido desenvolvidas para detectar textos nocivos, sem a necessidade de ter como base palavras ou frases específicas, a pesquisa demonstrou que, como Dias *et al.* colocam, elas estão "ainda distantes de conseguirem compreender o contexto ou identificar a intenção ou motivação do enunciador".[42] Conforme observado por Dias,[43] ainda que a correspondência de algoritmos (*hash-matching*) seja amplamente utilizada para identificar conteúdos de abuso sexual infantil, ela não é facilmente transposta para outros casos, como conteúdos extremistas, o qual "normalmente exige uma avaliação do contexto".

Nesse sentido, Keller observou que a decisão das plataformas de remover conteúdos extremistas islâmicos "imputará o ônus de forma sistemática e injusta a usuários inocentes da internet que porventura estejam se comunicando em árabe, discutindo a política do Oriente Médio ou falando sobre o islamismo".[44] Ela menciona a remoção de uma oração (em árabe)

publicada no Facebook por alegadamente ter violado os Padrões da Comunidade. A oração dizia "Deus, antes do fim deste dia sagrado, perdoe os nossos pecados, abençoe a nós e a nossos entes queridos nesta vida e na vida após a morte com a sua onipotente misericórdia".

Além disso, Dias *et al.*[45] constataram que essas tecnologias simplesmente não foram elaboradas para identificar linguagens como a da comunidade LGBTQIA+, cuja "zombaria de falta de educação" e o uso de termos como "sapatão", "bicha" e "traveco" constituem uma forma de resgatar o poder e preparar integrantes da comunidade para "lidar com a hostilidade". Dias *et al.* apresentam inúmeros relatos de ativistas LGBTQIA+ sobre remoção de conteúdo, como o banimento de uma mulher trans do Facebook após ela ter exibido uma fotografia de seu novo penteado, definindo a si mesma como "traveco".[46] Outro exemplo dado por Dias é uma pesquisa que revelou que tweets em inglês de afrodescendentes dos Estados Unidos têm duas vezes mais chances de serem considerados ofensivos em comparação com outros tweets, refletindo assim a infiltração de vieses raciais na tecnologia.[47] Dias *et al.* apontaram que os "efeitos de confundir um dialeto" precisam ser levados em consideração a fim de evitar preconceitos raciais na identificação de discurso de ódio.[48] Isso reflete a importância de contextualizar o discurso – algo que vai em desencontro com a elaboração e a aplicação de mecanismos automatizados e pode representar riscos para a participação *on-line* de grupos minoritários. Ademais, tais mecanismos carecem da capacidade de entender as nuances e o contexto da língua e da comunicação humana. Por exemplo, o YouTube removeu seis mil vídeos que documentavam o conflito na Síria.[49] Isso levou ao fechamento da Qasioun News Agency,[50] um grupo de mídia independente que relata crimes de guerra na Síria. Diversos vídeos foram sinalizados como inadequados por um sistema automatizado projetado para identificar conteúdos extremistas. Conforme observa Dias,[51] outras tecnologias de correspondência algorítmica, como o PhotoDNA, também parecem operar em um estado de "cegueira contextual", o que pode ser o motivo da remoção desses vídeos. Além disso, o Facebook baniu a palavra *kalar* em Mianmar, visto que radicais deram a essa palavra uma "conotação depreciativa" e a usaram para atacar o povo rohingya que vive no país. A palavra foi captada por mecanismos automatizados que apagaram publicações que poderiam tê-la utilizado em outro contexto ou com outro significado (incluindo *kalar oat*, que significa camelo). O resultado foi a remoção de publicações condenando movimentos fundamentalistas em Mianmar. Exemplo disso é a publicação a seguir, que apresenta a opinião de uma pessoa afirmando que o nacionalismo extremo e o fundamentalismo religioso são elementos negativos:

Fonte: Arquivo da autora.

Diante dos exemplos acima, os problemas em utilizar a IA para tratar alegações de discurso de ódio resultam não apenas na violação da liberdade de expressão devido a bloqueios excessivos, mas também em violações do direito à não discriminação.

6 • Conclusões

O Conselho Europeu propôs dez recomendações que podem ser adotadas para proteger direitos humanos no que diz respeito ao emprego de IA. Elas incluem, a título de exemplo, o estabelecimento de um marco legal para a realização de avaliações do impacto dos sistemas de IA existentes sobre os direitos humanos; a avaliação de sistemas de IA por meio de consultas públicas; a obrigação de Estados-Membros de promover a implementação de normas de direitos humanos em empresas privadas (como empresas de mídia social); uma supervisão transparente e independente dos sistemas de IA, com uma atenção especial a grupos desproporcionalmente afetados pela IA, como minorias étnicas e religiosas; a devida atenção aos direitos humanos, principalmente à liberdade de expressão; a regra de que a IA deve sempre permanecer sob controle humano e de que os Estados devem oferecer acesso efetivo à reparação para vítimas de violações de direitos humanos em decorrência do funcionamento da IA. O CE também menciona a promoção de um letramento sobre IA. Com relação a este último item, há espaço para a oferta de treinamento e capacitação na área de direitos humanos às pessoas direta ou indiretamente envolvidas na aplicação de sistemas de IA.[52]

Essas recomendações são de fato úteis para aprimorar o atual cenário de utilização de mecanismos automatizados para responder a discursos de ódio *on-line*. Todavia, as empresas

de mídias sociais devem se precaver quanto às questões estruturais originadas a partir da implementação dos mecanismos para remoção de discurso de ódio. Em primeiro lugar, é necessário destacar que, conforme observado por Llanso,[53] os problemas anteriormente mencionados não podem ser enfrentados por meio de uma IA mais sofisticada. Além disso, como Perel e Elink-Koren observam, "o processo de traduzir atos legais em código inevitavelmente incorpora escolhas particulares sobre como a lei é interpretada, o que pode ser afetado por uma série de considerações extrajudiciais, incluindo premissas profissionais conscientes ou inconscientes de desenvolvedores de programas, bem como diversos incentivos privados de empresas".[54] Embora os mecanismos automatizados possam auxiliar a moderação humana por meio da captação de possíveis discursos de ódio, eles não podem ser os únicos responsáveis pela remoção de tais discursos. Conjuntos de dados de treinamento enviesados, a ausência de dados relevantes e a falta de conceitualização do contexto e das nuances podem induzir a decisões equivocadas, tendo efeitos catastróficos na capacidade de grupos minoritários de operar de forma igualitária no mundo virtual.

NOTAS

1 • Jacob Mchangama *et al.*, "A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of Platformization," Justitia, novembro de 2021, acesso em 25 de novembro de 2022, https://futurefreespeech.com/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf.

2 • Alexandra Siegel *et al.*, "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath." Alexandra Siegel, 6 de março de 2019, acesso em 5 de janeiro de 2022, https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_et_al_election_hatespeech_qjps.pdf.

3 • Jacob Mchangama *et al.*, "A Framework of First Reference," Novembro de 2021.

4 • Jacob Mchangama e Natalie Alkiviadou, "The Digital Berlin Wall: How Germany Built a Prototype for Online Censorship." Euractiv, 8 de outubro de 2020, acesso em 4 de janeiro de 2022, https://www.euractiv.com/section/digital/opinion/the-digital-berlin-wall-how-germany-built-a-prototype-for-online-censorship/?fbclid=IwAR1fRPCtnP5ce_Glx77uaIB1sIS37BqqHdo-SliBiQWkYmGD3y7f8DaPOi4.

5 • "The Digital Services Act: ensuring a safe and accountable online environment," Comissão Europeia, 2022, acesso em 17 de outubro de 2022, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#documents.

6 • Emma J. Llansó, "No Amount of AI in Content Moderation Will Solve Filtering's Prior-Restraint Problem," *Big Data & Society* 7, no. 1 (2020).

7 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online," *Sexuality & Culture* 25 (2021): 700-732.

8 • Comitê de Especialistas do Conselho Europeu para o Desenvolvimento do Relatório de Direitos Humanos (2007) Capítulo IV, 123, § 4.

9 • Natalie Alkiviadou, "Regulating Hate Speech in the EU," in *Online Hate Speech in the EU: A Discourse Analytical Perspective*, 1ª ed., eds. Stavros Assimakopoulos, Fabienne H. Baider e Sharon Millar (Springer Cham, 2017).

10 • Recomendação 97 (20) do Comitê de Ministros

do Conselho Europeu sobre Discurso de Ódio.

11 • *Gűndűz vs. Turquia*, Petição no. 35071/97 (CEDH 4 de dezembro de 2003) § 40; *Erbakan vs. Turquia*, Petição no. 59405/00 (6 de julho de 2006) § 56.

12 • *Vejdeland e Outros vs. Suécia*, Petição no. 1813/07 (CEDH 9 de fevereiro de 2012) § 54.

13 • *Ibid.*

14 • *Ibid.* § 55.

15 • Fundamental Rights Agency, "Hate Speech and Hate Crimes against LGBT Persons" (2009) 1.

16 • Fundamental Rights Agency, "Homophobia and Discrimination on Grounds of Sexual Orientation and Gender Identity in the EU Member States: Part II - The Social Situation" (2009) 44.

17 • *The Observer e The Guardian vs. Reino Unido*, Petição no. 13585/88 (CEDH 26 de novembro de 1991) § 59.

18 • "Hate Speech," Meta Transparency Center, 2022, acesso em 25 de outubro de 2022, https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech.

19 • "Hate Speech," Meta Transparency Center, 2022, acesso em 2 de novembro de 2022, https://transparency.fb.com/policies/community-standards/hate-speech/#policy-details.

20 • "Promoting Hate Based on Identity or Vulnerability," Reddit, 2020, acesso em 2 de novembro de 2022, https://www.reddithelp.com/hc/en-us/articles/360045715951.

21 • "Hateful Conduct Policy," Twitter, 2016, acesso em 2 de janeiro de 2022, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

22 • "Hate Speech Policy," YouTube, 2019, acesso em 2 de janeiro de 2022, https://support.google.com/youtube/answer/2801939?hl=en.

23 • "Community Guidelines," TikTok, 2022, acesso em 2 de outubro de 2022, https://www.tiktok.com/community-guidelines?lang=en#38.

24 • Thiago Oliva Dias, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression," *Human Rights Law Review* 20, no. 4 (2020): 607-640.

25 • "How Technology Detects Violations," Meta Transparency Center, 19 de janeiro de 2022, acesso em 3 de novembro de 2022, https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/.

26 • "YouTube Community Guidelines enforcement," YouTube, 2022, acesso em 3 de novembro de 2022, https://transparencyreport.google.com/youtube-policy/removals.

27 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

28 • Natasha Duarte e Emma J. Llansó, "Mixed Messages? The Limits of Automated Social Media Content Analysis." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81 (2018): 106-106.

29 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

30 • Joch Cowls *et al.*, "Freedom of Expression in the Digital Public Sphere," AI and Platform Governance, 2020, acesso em 25 de novembro de 2022, https://doi.org/10.5281/zenodo.4292408.

31 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

32 • Emma J. Llansó, "No Amount of AI...," 2020.

33 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation and Freedom of Expression." Transatlantic Working Group, 2020, acesso em 23 de novembro de 2022, https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

34 • "Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications," Council of Europe, DGI (2017) 12, 2017, acesso em 23 de novembro de 2022, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, 18.

35 • *Ibid.* 21.

36 • Robert Gorwa *et al.*, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7, no. 1 (2020).

37 • Emma Llanso *et al.*, "Artificial Intelligence,

Content Moderation...," 2020.

38 • "Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression," OHCHR, 13 de junho de 2018, acesso em 10 de novembro de 2022, https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

39 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

40 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation...," 2020.

41 • Natasha Duarte e Emma J. Llansó, "Mixed Messages?...," 2018.

42 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

43 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

44 • Daphne Keller, "Internet Platforms: Observations on Speech, Danger and Money," *Hoover Institution's Aegis Paper Series*, no. 1807 (2018).

45 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

46 • *Ibid.*

47 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

48 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

49 • "YouTube 'made wrong call' on Syria videos'," BBC News, 23 de agosto de 2017, acesso em outubro de 2022, https://www.bbc.com/news/technology-41023234.

50 • *Ibid.*

51 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

52 • "Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights," Conselho Europeu, 2019, acesso em 23 de novembro de 2022, https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

53 • Emma J. Llansó, "No Amount of AI...," 2020.

54 • Maayan Perel e Niva Elkin-Koren, "Accountability in Algorithmic Copyright Enforcement," *19 Stanford Technology Law Review* 473 (2016), acesso em 25 de novembro de 2022, https://law.stanford.edu/wp-content/uploads/2016/10/Accountability-in-Algorithmic-Copyright-Enforcement.pdf.

NATALIE ALKIVIADOU – *Chipre/Dinamarca*
Natalie Alkiviadou é Pesquisadora Sênior da Justitia (Dinamarca). Seus temas de interesses são liberdade de expressão, extrema-direita, discursos de ódio e crimes de ódio. Ela já publicou três monografias e uma série de artigos revisados por pares. Alkiviadou é integrante do Centro de Direito da Sociedade da Informação da Università degli Studi di Milano.

contato: *natalie@justitia-int.org*

Recebido em setembro de 2022.
Original em inglês. Traduzido por Naiade Rufino.