A DESINFORMAÇÃO NAS PLATAFORMAS DIGITAIS E A "POLÍTICA DE INFORMAÇÕES ENGANOSAS SOBRE A COVID-19", DO TWITTER: DESAFIOS E PERSPECTIVAS DA MODERAÇÃO DE CONTEÚDO

Pillar Cornelli Crestani*
Rafael Santos de Oliveira**

RESUMO

O presente artigo objetiva verificar quais as estratégias adotadas pelo *Twitter*, para combater a desinformação que circula em seu domínio, no contexto da pandemia do Novo Coronavírus, bem como os desafios e perspectivas da moderação de conteúdo da plataforma. Diante deste problema, explicitado pelo método de abordagem dedutivo, a pesquisa parte de uma situação ampla, demonstrada pelo fenômeno da desinformação virtual, com ênfase na pandemia da Covid-19, encaminhando-se para a verificação de um caso específico, demonstrado por meio da análise da "Política de informações enganosas sobre a Covid-19", do *Twitter*. Associado a esse referencial metodológico, o presente estudo utilizou o método de procedimento monográfico, combinando as técnicas de pesquisa bibliográfica e documental, pois foi analisada a "Política de informações enganosas sobre a Covid-19", no site do *Twitter*, bem como os estudos de especialistas da área do Direito Digital e da tecnologia. Por fim, concluiu-se que o *Twitter* proíbe, como regra geral, a publicação de conteúdos falsos sobre a Covid-19 que possam causar danos aos indivíduos. Além disso, apurou-se que, como forma de sanção ao descumprimento de suas diretrizes, a plataforma estabelece determinadas categorias de intervenções a serem aplicadas às postagens violadoras de sua política, a saber: a exclusão do *tweet*; a marcação do conteúdo violador; e o bloqueio ou a suspensão permanente da conta do usuário do *Twitter* – medidas que devem ser utilizadas com a devida transparência, a fim de evitar a violação de direitos fundamentais dos usuários e da coletividade.

Palavras-chave: COVID-19; desinformação virtual; direito informacional; moderação de conteúdo; *Twitter*.

Data de submissão: 01/09/2022
Data de aprovação: 20/03/2023

* Mestra em Direito pela Universidade Federal de Santa Maria (UFSM).

** Doutor em Direito pela Universidade Federal de Santa Catarina (UFSC).

DISINFORMATION ON DIGITAL PLATFORMS AND TWITTER'S "POLICY OF MISLEADING INFORMATION ABOUT COVID-19": CHALLENGES AND PERSPECTIVES OF CONTENT MODERATION

Pillar Cornelli Crestani
Rafael Santos de Oliveira

ABSTRACT

This present article objectives to verify which are the strategies adopted by Twitter to oppose misinformation on its platform, in the context of the Covid-19 pandemic. From the formulation of this problem, expressed by the deductive approach method, this research departed of a broad situation, demonstrated by the virtual disinformation phenomenon, especially in the Covid-19 pandemic, going to deductive verify a specific case, that is demonstrated by "Twitter's Covid-19 Misleading Information Policy". Allied to this methodological reference, the present study used the procedure monographic method, combining bibliographic and documentary research techniques, because was analyzed the "Twitter's Covid-19 Misleading Information Policy", on the Twitter website, and the studies of specialists in the field of Digital Law and technology. Based on the study, it was concluded that Twitter prohibits the posting of false content about Covid-19 that could harm people and, as a form of sanction for non-compliance with the guidelines, the platform establishes certain categories of intervention measures to be applied to posts that violate its policy, like: the exclusion of the tweet; marking offending content; and blocking or permanently suspending the user's Twitter account. These measures, however, must be used with due transparency, in order to avoid violating the fundamental rights of users and the community.

Keywords: COVID-19; virtual disinformation; right to information; content moderation; Twitter.

Date of submission: 01/09/2022
Date of approval: 20/03/2023

INTRODUÇÃO

As plataformas de conteúdo gerado pelos internautas – também chamadas de "provedores de aplicação" – representam uma grande evolução no âmbito das comunicações, em razão de terem oportunizado, aos indivíduos, a ampliação da faculdade de manifestar as próprias ideias e opiniões. Por conta disso, entretanto, essas plataformas acabaram se tornando um ambiente marcado por uma intensa desordem informacional (caracterizada, sobretudo, pela difusão de conteúdos impróprios, ofensivos a direitos de terceiros e desinformativos), especialmente, pela falsa concepção de que a internet propicia o exercício absoluto da liberdade de expressão, desprovido de qualquer limitação ou consequência de ordem legal.

Nesse sentido, há que se destacar que houve um agravamento dessa conjuntura durante a pandemia do Novo Coronavírus (SARS-CoV-2), o que motivou a Organização Mundial da Saúde (OMS) a declarar a existência de uma "infodemia" – expressão que designa a "pandemia de informação" enfrentada no contexto da crise sanitária global vigente (Organização Pan-Americana da Saúde, 2020, p. 2) – tendo em vista, sobretudo, a desmedida propagação de conteúdos desinformativos sobre a Covid-19, principalmente, nas plataformas digitais.

Diante disso, não se pode deixar de evidenciar que a desinformação é potencialmente prejudicial nesse cenário pandêmico, pois compromete o enfrentamento ao vírus, em termos de saúde pública, e expõe a vida da população a riscos graves. Além disso, destaca-se que esse fenômeno também viola o direito informacional dos indivíduos – os quais possuem a prerrogativa de acesso a informações verídicas, tanto por parte dos meios jornalísticos, como também por parte das fontes informais de comunicação, nas quais se enquadram as plataformas digitais.

Nessa perspectiva, evidencia-se que grande parte das plataformas digitais de conteúdo gerado pelos internautas – cientes da nocividade da desinformação sobre a Covid-19 e de que o seu domínio constitui uma das principais vias de propagação dos conteúdos desinformativos sobre a pandemia – adotaram determinadas medidas visando a contribuir com o enfrentamento da crise sanitária, no âmbito virtual. A partir disso, levando em consideração que o *Twitter*, apesar de não figurar no rol das plataformas com maior número de usuários no Brasil (Volpato, 2022), constitui um *locus* relevante de debate público, representando um espaço multimídia, onde são travadas, de forma instantânea, as principais discussões sobre os assuntos do momento, questiona-se: quais as estratégias adotadas pelo *Twitter*, para combater a desinformação sobre a Covid-19, em seu domínio, e quais os desafios e as perspectivas relacionados ao sistema de moderação de conteúdos da plataforma?

A partir desse problema de pesquisa, como objetivo enfrentado pelo presente trabalho, busca-se verificar quais as medidas tomadas pela plataforma em questão, com vistas a confrontar os conteúdos desinformativos que circulam em seu domínio. Para tanto, emprega-se o método de abordagem dedutivo, eis que a pesquisa parte de uma situação ampla, explicitada pelo fenômeno da desinformação virtual, com destaque para a pandemia da Covid-19, encaminhando-se para a verificação de um caso específico, demonstrado por meio da análise da "Política de informações enganosas sobre a Covid-19", do *Twitter*.

Aliado a esse referencial metodológico, o presente estudo recorre ao método monográfico, combinando as técnicas de pesquisa bibliográfica e documental, pelo fato de ter sido analisada a "Política de informações enganosas sobre a Covid-19", no site do *Twitter*, bem como os estudos de especialistas da área do Direito Digital e da tecnologia – tudo isso com vistas a responder ao problema proposto por esta produção.

Nessa perspectiva, expõe-se que a aplicação do referido método resultou na divisão do artigo em duas partes: primeiramente, contextualiza-se a conjuntura da desinformação virtual, com ênfase no contexto da crise sanitária global decorrente do Novo Coronavírus. Por conseguinte, procede-se a uma análise da "Política de informações enganosas sobre a Covid-19", do *Twitter*, a fim de verificar quais as medidas tomadas pela plataforma em questão, para combater os conteúdos desinformativos que circulam em seu domínio, bem como os desafios e perspectivas relacionados ao seu sistema de moderação.

Por conseguinte, destaca-se a relevância da discussão proposta pelo presente trabalho, tendo em vista que a desinformação virtual consiste em um fenômeno global, que vem sendo intensificado nos últimos tempos, provocando inúmeros prejuízos às pessoas e, também, às democracias. E, de forma específica, no contexto da pandemia da Covid-19, reitera-se que a desinformação compromete o enfrentamento da crise sanitária, expondo a saúde coletiva a risco e violando o direito informacional dos indivíduos.

Nesse sentido, ainda, entende-se pertinente averiguar o posicionamento das plataformas de conteúdo gerado pelos internautas – como é o caso do *Twitter* – no que tange à desinformação relacionada a todos os aspectos que envolvem a pandemia do Novo Coronavírus. Isso porque essas plataformas, enquanto mediadoras do discurso público e "tendo em vista o impacto social e político que circunda suas atividades" (Hartmann; Iunes, 2020, p. 395), devem contribuir para resguardar a sua comunidade de informações enganosas que possam vir a prejudicar a saúde de seus usuários no mundo real *offline*, adotando, portanto, medidas de combate à desinformação pandêmica, por intermédio de seu sistema de moderação de conteúdos, conforme será abordado na sequência.

1 A DESINFORMAÇÃO NAS PLATAFORMAS DIGITAIS, NO CONTEXTO DA PANDEMIA DA COVID-19

A evolução das tecnologias de informação proporcionou significativos avanços em todos os âmbitos da sociedade em rede, ampliando a capacidade comunicacional dos indivíduos, especialmente, no meio virtual (Castells, 2015, p. 58). Essa nova estrutura permitiu, "pela primeira vez, a comunicação de muitos com muitos" (Castells, 2003, p. 8), além de ter propiciado, aos indivíduos, a oportunidade de protagonizar a produção de conteúdos e compartilhá-los no ciberespaço (Levy, 2002, p. 52).

Destaca-se que todas essas vantagens foram viabilizadas, efetivamente, a partir do surgimento das plataformas de conteúdo gerado pelos internautas[1] – a exemplo de *Twitter*, *Facebook*, *YouTube*. Por meio delas, os seus usuários podem "produzir conteúdo para acesso dos demais, sem haver uma curadoria prévia ou contrato comercial entre os produtores de conteúdo e a plataforma que o disponibiliza" (Rodrigues; Kurtz, 2020, p. 17).

Por conta disso, entretanto, essas plataformas acabaram se tornando um terreno fértil para a propagação de desinformação, especialmente, pela ideia equivocada de que o ambiente virtual propicia o exercício absoluto da liberdade de expressão, descompromissado em relação ao respeito aos direitos de terceiros e à veracidade dos conteúdos publicados nas plataformas. Desse modo, evidencia-se que toda essa conjuntura contribuiu para gerar uma "sobrecarga de informação" no espaço virtual, acarretando a violação do direito informacional da coletividade, bem como de outras garantias fundamentais dos indivíduos (Alves; Maciel, 2020, p. 149).

Nessa perspectiva, enfatiza-se que existe uma grande amplitude de conceitos que dão conta de designar o fenômeno da desinformação, não existindo, portanto, um significado unânime, na literatura acadêmica ou no discurso jornalístico, a seu respeito (Ribeiro; Ortellado, 2018, p. 72). O fato é que, frequentemente, a desinformação acaba sendo definida, de forma genérica, pelo termo "fake news" ("notícias falsas", em Português) – o que representa um grande equívoco. Isso porque, nem sempre, os conteúdos desinformativos consistem em "notícias" e, não necessariamente, são integralmente falsos, pois existe a possibilidade de haver a distorção – de modo involuntário ou intencional, com o propósito específico de enganar – de informações que, de fato, são verídicas (Wardle, 2019).

Todavia, destaca-se que, no presente trabalho, será adotada a seguinte definição para referir-se aos conteúdos desinformativos: "informação verificável como falsa ou enganosa que tem o potencial de causar dano ao público, como enfraquecer a democracia ou prejudicar a saúde pública" (O Brasil, 2020). Trata-se, portanto, de um conceito que se mostra adequado na atual conjuntura informacional da sociedade em rede, visto que a desinformação vem provocando sérios impactos em termos democráticos e de saúde coletiva, especialmente, no contexto da pandemia da Covid-19.

Por conseguinte, não se pode deixar de comentar, brevemente, que a desinformação virtual contempla determinadas categorias, considerando o seu potencial de gerar danos no mundo *offline*, de acordo com as definições apresentadas por Wardle (2019a)[2]. Dentre elas, estão incluídas, por exemplo, as sátiras e paródias, que demonstram baixo potencial danoso, pois não possuem o

[1] Há que se destacar que "plataformas de conteúdo gerado por uuário (sic) são uma gama ampla de comunidades e serviços, que incluem tanto redes sociais quanto ferramentas de compartilhamento/disponibilização de vídeos, comentários etc. Diferenciam-se, ainda, de outros tipos de plataforma online, como as de streaming, de compartilhamento de bens e serviços, ou de notícias, por exemplo, pelo caráter de autonomia de cada usuário (Rodrigues; Kurtz, 2020, p. 17).

[2] Claire Wardle, pesquisadora do *First Draft*, propõe sete categorias para classificar a desordem informacional, baseadas no grau de prejuízos que os conteúdos desinformativos podem ocasionar, do menor para o maior, a saber: sátira ou paródia; falsa conexão; conteúdo enganoso; falso contexto; conteúdo impostor; conteúdo manipulado; e conteúdo fabricado (Wardle, 2019a).

propósito de enganar, mas sim, o de expressar alguma crítica por meio do humor – o que pode acabar gerando algum equívoco, caso a interpretação do interlocutor seja distinta do sentido expresso (pretendido) pelo criador do conteúdo. Por outro lado, existem os chamados "conteúdos fabricados", que, por sua vez, são integralmente falsos, pois são projetados no intuito de gerar engano e, consequentemente, prejudicar os consumidores dessas informações.

Prosseguindo na temática proposta pelo presente trabalho, entende-se imprescindível destacar que o fenômeno da desinformação virtual decorre da atuação de inúmeros atores, motivados por razões diversas. Nessa perspectiva, evidencia-se que os próprios internautas podem ser considerados grandes difusores de conteúdos desinformativos nas plataformas digitais de conteúdo gerado pelo usuário, pelo fato de não se interessarem em efetuar a checagem das informações que recebem e compartilham[3] ou, simplesmente, por publicarem conteúdos (muitas vezes, de sua própria autoria) baseados em suas convicções pessoais[4], desprovidos, portanto, de qualquer compromisso com a veracidade das informações propagadas (Manjoo, 2008; Spinelli; Santos, 2018, p. 763).

Entretanto, no ecossistema da desinformação virtual, sobressai-se o protagonismo de agentes maliciosos dedicados, especificamente, à criação de conteúdos de teor inverídico ou distorcido, impulsionados por motivos políticos, ideológicos e econômicos (Teffé, 2018). Nesse sentido, expõe-se que essas informações falsas podem ser utilizadas no intuito de prejudicar "o outro lado", o "rival", ou, ainda, de implantar determinadas ideias à coletividade. E, além disso, ressalta-se que a desinformação consiste em uma "indústria" bastante lucrativa, pois os conteúdos enganosos – normalmente, de cunho apelativo e sensacionalista – recebem alto engajamento nas redes, atraindo milhares de cliques, os quais são monetizados, em decorrência das publicidades às quais estão atrelados, nos *sites* (Tandoc *et al.*, 2018, p. 2 *apud* Brites; Amaral; Catarino, 2018, p. 86).

Tudo isso se torna possível, dentre outras razões, especialmente, pelo fato de a desinformação afetar diretamente o psicológico dos indivíduos. Destaca-se que a maioria dos conteúdos enganosos é planejada a partir da utilização de técnicas de persuasão e de manipulação das emoções das pessoas, especialmente, em contextos de crises sanitárias, como a pandemia da Covid-19. Nesses casos, os agentes desinformadores se aproveitam do sentimento de medo e de incerteza – os quais são inerentes aos seres humanos nesses cenários extraordinários, marcados por vulnerabilidades de todas as ordens – para atingir os seus objetivos e interesses: propagar ideias próprias ou, apenas, ludibriar a população (Taylor, 2019, p. 65).

Como exemplo disso, é possível citar a disseminação de teorias da conspiração e boatos a respeito da vacinação, o que acaba gerando um cenário

---

[3] Essa questão pode ser explicada por meio da ideia da "avareza cognitiva", pois "preferimos usar maneiras mais simples e fáceis de resolver problemas do que aquelas que exigem mais reflexão e esforço. Evoluímos para usar o mínimo de esforço mental possível" (Shane, 2020, p. 3, tradução nossa).

[4] Nessa perspectiva, convém destacar a questão do "viés de confirmação", uma característica do ecossistema da desinformação que gera uma "tendência a acreditar em informações que confirmam suas crenças existentes e rejeitar informações que as contradizem" (Shane, 2020, p. 7, tradução nossa).

de hesitação vacinal, em decorrência do medo, por parte de alguns indivíduos, de serem acometidos por supostas reações adversas prejudiciais ao organismo, ao receberem algum imunizante. Nesse sentido, entende-se que a desinformação é gravemente prejudicial, pois o desencorajamento à imunização acaba favorecendo a propagação das doenças combatidas por meio das vacinas, colocando em risco, portanto, a saúde da coletividade (Taylor, 2019, p. 87).

Por conseguinte, também não se pode deixar de mencionar que a propagação da desinformação é viabilizada pela arquitetura das plataformas digitais de conteúdo gerado pelo usuário (e da própria internet, em si). Isso porque os conteúdos são difundidos, no meio virtual, em um curto lapso de tempo, sendo alcançados por milhares de usuários das redes, que, por sua vez, impulsionam essas postagens aos seus contatos, em proporções globais[5]. Desse modo, levando em consideração as características dos discursos desinformativos mencionadas anteriormente (o tom apelativo e sensacionalista), os conteúdos enganosos acabam "viralizando" e obtendo engajamento em dimensões inimagináveis.

Outro aspecto relevante, no ecossistema da desinformação virtual, são as chamadas "bolhas", que são constituídas a partir dos filtros dos algoritmos das plataformas digitais. A atividade algorítmica de filtragem reúne e interliga conteúdos que possuem o mesmo padrão e as mesmas características, fazendo, por exemplo, com que o *feed* de um usuário que consome e gera engajamento a postagens desinformativas seja retroalimentado por esse tipo de conteúdo (Pariser, 2012). Explicita-se que essa questão dos filtros-bolha é extremamente problemática, à medida que acentua as polarizações na sociedade, impedindo a pluralidade e o confronto de ideias – o que é essencial, especialmente, no âmbito da democracia.

Entretanto, destaca-se que essa atividade algorítmica de filtragem é essencial à atividade das plataformas digitais, tendo em vista que a coleta de dados de seus usuários permite que o sistema de inteligência artificial segmente os conteúdos que serão direcionados, de modo personalizado, a cada um dos perfis dos internautas. Por essa razão, quanto mais um indivíduo interagir (curtir, comentar e replicar) com postagens enganosas, mais esse tipo de conteúdo será recomendado a ele, fazendo com que esteja inserido em uma bolha desinformativa (Pariser, 2012). E todos esses aspectos estão interligados à economia da atenção (Wu, 2016)[6] e ao capitalismo de vigilância (Zuboff, 2021)[7], à medida que o lucro

[5] A título de observação, há que se ter em vista que a propagação da desinformação também se dá, em grande parte, por meio de contas falsas, nas redes sociais, movimentadas por robôs, que são criadas com o intuito exclusivo de disseminar conteúdos enganosos e prejudiciais. O combate a esses perfis automatizados – que configuram verdadeiros "exércitos digitais" – representam um grande desafio às plataformas, em razão da dificuldade em serem identificados, para fins de sua desativação, e, também, da rapidez com que são criados e replicados, formando, assim, um ciclo permanente (Avaaz, 2019).

[6] Em síntese, destaca-se que a ideia de "economia da atenção" proposta por Tim Wu (2016) está concentrada no fato de que a atenção das pessoas é mercantilizada, por meio do consumo das publicidades atreladas aos meios de comunicação e, nos últimos tempos, às plataformas digitais, os quais constituem os "mercadores da atenção" (Wu, 2016).

[7] O "capitalismo de vigilância" é marcado pela coleta massiva de dados dos indivíduos, pelas próprias plataformas digitais, pela qual é possível estabelecer previsões a respeito do comportamento dos usuários, com a finalidade de lhes direcionar anúncios personalizados de produtos e serviços (Zuboff, 2021).

obtido pelas plataformas de conteúdo gerado pelos usuários é proporcional ao tempo despendido pelos internautas nesses sítios.

Por fim, convém ressaltar que o fenômeno da desinformação também é facilitado pelo anonimato – outra característica da arquitetura da rede – o que acaba estimulando, de certa forma, a circulação de conteúdos enganosos, ofensivos e violadores de direitos humanos. Tudo isso porque a identificação e a localização dos internautas infratores, por vezes, torna-se desafiadora, para fins de responsabilização dos atos ilícitos cometidos, gerando a falsa aparência de que a internet constitui um ambiente desprovido de normas regulamentadoras da conduta de seus usuários (Alvez; Maciel, 2020, p. 149).

Dando sequência à temática proposta pelo presente trabalho, destaca-se que toda essa conjuntura da desinformação virtual foi profundamente agravada durante a pandemia da Covid-19, considerada a "maior crise sanitária mundial da nossa época" (OMS, 2020). Nesse sentido, haja vista o intenso fluxo de conteúdos desinformativos, no ciberespaço, a respeito de todos os âmbitos relacionados ao contexto pandêmico, a Organização Mundial da Saúde (OMS) declarou a existência de uma "infodemia" – expressão que, justamente, designa:

> [...] um grande aumento no volume de informações associadas a um assunto específico, que podem se multiplicar exponencialmente em pouco tempo devido a um evento específico, como a pandemia atual. Nessa situação, surgem rumores e desinformação, além da manipulação de informações com intenção duvidosa. Na era da informação, esse fenômeno é amplificado pelas redes sociais e se alastra mais rapidamente, como um vírus (Organização Pan-Americana da Saúde, 2020, p. 2).

A partir desse conceito e levando em consideração todos os fatores que facilitam a propagação dos conteúdos desinformativos no meio virtual, comentados anteriormente, não se pode deixar de evidenciar que a desinformação agrava potencialmente o cenário da pandemia. Isso porque obstaculiza o acesso, por parte do público, a informações verídicas provenientes da comunidade científica e das autoridades de saúde, dificultando a prática de condutas necessárias ao enfrentamento da crise sanitária, como o uso de máscaras de proteção, o distanciamento social e a adesão à vacinação, conforme comentado anteriormente (Organização Pan-Americana da Saúde, 2020, p. 3).

Do mesmo modo, afirma-se que as informações enganosas também podem influenciar os indivíduos a adotarem medidas prejudiciais à sua saúde, como é o caso dos conteúdos que recomendam falsas curas à Covid-19, bem como tratamentos inadequados, desprovidos de qualquer embasamento científico. Nesse cenário, destaca-se que a desinformação expõe a vida dos indivíduos a risco, além de violar o direito informacional da coletividade[8] – a qual possui a prerrogativa de receber

[8] Entende-se conveniente destacar o conceito do direito à informação, que, nas palavras de Anderson Schreiber, "caracteriza-se como direito de receber, acessar ou difundir informações, sendo relevante, nesse aspecto, o caráter de veracidade objetiva da informação transmitida", não compreendendo uma prerrogativa exclusiva de jornalistas, mas, também, de todos os indivíduos (Schreiber, 2018, p. 65).

informações adequadas e verídicas, não só provenientes das mídias tradicionais, mas, também, das fontes informais de comunicação, dentre as quais estão incluídas as plataformas digitais de conteúdo gerado pelo usuário.

Diante disso, comenta-se que essas plataformas de conteúdo gerado pelos internautas – também chamadas de "provedores de aplicação", de acordo com a definição estabelecida pelo artigo 5º, inciso VII, do Marco Civil da Internet (Brasil, 2014) – cientes de que grande parte da desinformação virtual a respeito da Covid-19 circula em seu domínio, adotaram determinadas medidas para combater a "infodemia" e, assim, contribuir para o enfrentamento da crise sanitária vigente, no mundo *offline* (Agrela, 2020).

Nessa perspectiva, esclarece-se que, por questões de viabilidade da presente pesquisa, optou-se por verificar quais as estratégias adotadas pelo *Twitter*[9], para combater a desinformação sobre a Covid-19 que é propagada em seu domínio. A opção pela referida plataforma digital deu-se pelo fato de esta representar um espaço onde circula, de forma instantânea, uma vasta quantidade de informações sobre os acontecimentos do momento e onde são travadas discussões sobre os mais variados assuntos (Comm, 2009). Apesar de não constituir a plataforma com o maior número de usuários no Brasil[10], o público do *Twitter* vem crescendo desde 2020, sendo utilizado, "principalmente como segunda tela em que os usuários comentam e debatem o que estão assistindo na TV, postando comentários sobre noticiários, reality shows, jogos de futebol e outros programas" (Volpato, 2022).

Além disso, a opção pela referida plataforma digital também se deu pelo fato de, no mês de março de 2020, quando foi decretada a pandemia, pela OMS, "cerca de 550 milhões de tuítes continham os termos coronavirus, corona virus, covid19, covid-19, covid_19 ou pandemic [pandemia]" (Organização Pan-Americana da Saúde, 2020, p. 2). Assim, diante do intenso fluxo informacional acerca da crise sanitária provocada pelo Novo Coronavírus – levando-se em consideração, ainda, todos os conteúdos desinformativos incluídos nesse conjunto – na próxima seção, serão abordadas as estratégias adotadas pelo *Twitter*, para combater a desinformação sobre a Covid-19, em seu domínio, bem como os desafios e perspectivas da moderação de conteúdo da plataforma.

2 A "POLÍTICA DE INFORMAÇÕES ENGANOSAS SOBRE A COVID-19", DO TWITTER, E OS DESAFIOS E PERSPECTIVAS DA MODERAÇÃO DE CONTEÚDO DA PLATAFORMA

Conforme já comentado no primeiro capítulo, ciente de seu papel enquanto promotor do discurso coletivo e de vetor de conteúdos desinformativos a respeito da pandemia do Novo Coronavírus, o *Twitter* aderiu a estratégias para combater a

[9] Entende-se oportuno evidenciar, em síntese, que o *Twitter* constitui uma "rede social de microblogs, onde os usuários podem escrever mensagens de até 140 caracteres. Os usuários são identificados por @nome_do_usuário e os assuntos podem ser categorizados por hashtags (#)" (TIC Domicílios, 2019, p. 373).

[10] No início de 2022, o *Twitter* possuía apenas 19 milhões de usuários ativos no Brasil – o que configura um número reduzido, se comparado às plataformas mais populares do país, como é o caso do *WhatsApp*, que é utilizado por 165 milhões de usuários, e do *YouTube*, que é consumido por 138 milhões de internautas (Volpato, 2022).

circulação de desinformação sobre a Covid-19 em seu domínio, seguindo, então, a tendência das demais plataformas digitais. Nessa perspectiva, explicita-se, inicialmente, que as iniciativas adotadas pela referida plataforma estão reunidas na "Política de informações enganosas sobre a Covid-19", as quais compõem uma seção especial, na "Central de Ajuda" do *site*.

Esclarece-se que essa política compreende um conjunto de diretrizes que orientam a conduta que deve ser incorporada pelos membros do *Twitter*, em seu domínio, dispondo sobre os conteúdos que não são permitidos de serem postados na plataforma, bem como as consequências a serem aplicadas em caso de violação dessas determinações. Nesse sentido, há que se destacar que a regra geral consiste na proibição do compartilhamento de "informações falsas ou enganosas sobre a Covid-19 que possam causar danos", sob a justificativa de proteger a coletividade dos efeitos nocivos da desinformação (Política, 2021).

Entretanto, não se pode deixar de ressaltar que a "Política de informações enganosas sobre a Covid-19", do *Twitter*, junto dos Termos de Uso, está inserida no contexto do sistema de moderação de conteúdos da plataforma, o qual constitui um "mecanismo de governança que estrutura a participação em uma comunidade para facilitar a cooperação e evitar abusos" (Grimmelmann, 2015, p. 47)[11]. Sem o intuito de esgotar a temática, que é complexa e suscita inúmeras discussões, comenta-se, brevemente, que:

> A regulação sobre plataformas digitais está inserida em um contexto de governança multissetorial, havendo três modelos regulatórios principais que podem ser adotados, comumente mencionados pela literatura: autorregulação, corregulação (ou autorregulação regulada) e regulação por comando e controle (top-down) (Almeida, 2022, p. 116).

De acordo com Gorwa (2019, p. 11), as plataformas digitais estão assentadas no modelo da autorregulação, ou seja, enquanto entes privados, elas próprias instituem as suas regras e dirimem conflitos com os usuários ou entre eles, sem a interferência de um agente externo, como o Estado, por exemplo, nesses processos. Ou seja: a moderação de conteúdos – intrínseca à regulação autônoma – constitui uma atividade de praxe, na maioria das plataformas digitais de conteúdo gerado pelos usuários, com o objetivo de manter o seu domínio o mais "saudável" possível para os internautas, estabelecendo, portanto, limites à conduta destes, no âmbito *online*. Diante disso, as plataformas estabelecem regras, determinando quais categorias de publicações são proibidas e permitidas de serem veiculadas, do contrário, esses espaços virtuais acabariam se assemelhando à *dark web*, tendo em vista a circulação de postagens sensíveis, impróprias e violadoras de direitos humanos – as quais não se coadunam com a proposta da internet convencional, a *surface web* (Kurtz; Do Carmo; Vieira, 2021).

[11] Vale destacar que "governança e regulação são conceitos distintos. Enquanto a governança apresenta-se como uma estrutura complexa em rede que acomoda diferentes partes interessadas, com diferentes interesses, coordenadas de maneira colaborativa por diversos instrumentos, que tanto podem ter origem pública, como particular. A regulação é um mecanismo da governança, que objetiva conformar o comportamento de um determinado grupo, com consequências imprevisíveis para a governança" (Couto, 2022, p. 210).

Nessa perspectiva, evidencia-se que a principal motivação para as plataformas digitais exercerem a atividade moderadora é a questão comercial, pelo fato de ser necessário manter um ambiente atrativo para os anunciantes – os quais desejam não estar atrelados a conteúdos inapropriados, nas redes, pois, do contrário, comprometeria a sua reputação no mercado (Rodrigues; Kurtz, 2020). Também não se pode deixar de comentar que as regras que norteiam o sistema de moderação de conteúdos das plataformas – no caso do *Twitter*, os Termos de Uso e a "Política de informações enganosas sobre a Covid-19" – constituem um contrato de adesão entre a empresa e o usuário, com o qual este concorda integralmente, sem a possibilidade de discutir as cláusulas, ao registrar a sua conta na plataforma (Rodrigues; Kurtz, 2020, p. 27).

Por conseguinte, expõe-se que o sistema de moderação de conteúdos das plataformas digitais pode atuar de modo misto: por meio da inteligência artificial (algoritmos), que detectam postagens violadoras dos Termos de Uso e das políticas da empresa, aplicando as devidas medidas interventivas em relação ao conteúdo, quando cabível no caso; e por meio dos moderadores humanos, que efetuam uma análise mais aprofundada na publicação detectada pela filtragem algorítmica da plataforma, buscando verificar o contexto do *post*, e, com base nas regras preestabelecidas, optar pela aplicação ou não das intervenções adotadas pela empresa (Rodrigues; Kurtz, 2020, p. 16-18).

No que tange às medidas interventivas, esclarece-se que, em geral, as plataformas aplicam os seguintes métodos de moderação: remoção; indisponibilização (*geoblocking*); restrição ("desmonetização"); sinalização; ranqueamento (*shadowban*); suspensão e banimento de contas (Kurtz; Do Carmo; Vieira, 2021, p. 14-16). Destaca-se que algumas dessas alternativas moderatórias estão incluídas na "Política de informações enganosas sobre a Covid-19", do *Twitter*, conforme será visto na sequência.

Por fim, é necessário frisar que a regulação das plataformas, mais especificamente, a atividade de moderação de conteúdos, é absolutamente necessária, para resguardar os seus usuários e o público em geral, bem como para atrair novos membros, anunciantes e parceiros. Desse modo, apesar de constituírem espaços privados e "abertos", as plataformas não são neutras e, tampouco, desprovidas de regras – conforme preconizado pelo senso comum (Gillespie, 2018, p. 5).

Feitas essas considerações a respeito do sistema de moderação de conteúdos das plataformas digitais, para melhor compreensão das ideias a serem abordadas neste segundo capítulo, expõe-se que a "Política de informações enganosas sobre a Covid-19", do *Twitter*, aborda quatro categorias de conteúdos desinformativos proibidos de serem abordados no *Twitter*, as quais serão comentadas na sequência. A primeira delas refere-se a informações falsas ou enganosas sobre a natureza do vírus; sobre as formas de transmissão da doença; sobre a suscetibilidade de determinados indivíduos ao vírus; sobre os sintomas da doença; bem como orientações de automedicação e discursos conspiratórios a respeito das vacinas contra a Covid-19 (Política, 2021).

Por conseguinte, a segunda categoria engloba informações falsas ou enganosas sobre a eficácia e/ou segurança de medidas de prevenção, tratamentos ou outras precauções para mitigar ou tratar a Covid-19, incluindo-se aqueles que

não possuem a aprovação das autoridades sanitárias. Nesse conjunto, também consta a proibição de publicação de conteúdos que explanem reações adversas equivocadas em relação ao recebimento das vacinas contra o Novo Coronavírus, bem como manifestações envolvendo a ideia de que os imunizantes constituem "uma tentativa deliberada de causar danos ou controlar populações" (Política, 2021).

A terceira categoria de postagens vedadas pelo *Twitter* compreende informações falsas ou enganosas sobre regulamentações oficiais, restrições ou isenções relacionadas a orientações de saúde. Nela são elencados, por exemplo, os conteúdos questionando a segurança e a eficácia das máscaras de proteção e das demais medidas sanitárias para conter a disseminação do Novo Coronavírus. Nesse tópico, também está incluída a proibição de *tweets* enganosos a respeito do contexto que envolve o desenvolvimento, a testagem e a produção dos imunizantes contra a Covid-19 (Política, 2021) – evidenciando que, pelo menos, na teoria, a desinformação a respeito da vacinação representa, igualmente, uma preocupação concreta, por parte da plataforma, a qual também apresenta uma seção especial para tratar, exclusivamente, sobre essa questão.

Por fim, a quarta categoria de conteúdos proibidos de serem abordados no *Twitter* diz respeito a informações falsas ou enganosas sobre a prevalência do vírus, bem como sobre o risco de infecção ou morte, elencando, de modo exemplificativo, postagens contendo estatísticas inverídicas sobre o número de casos, de hospitalizações, de mortes e de outras variáveis relacionadas à pandemia, tal qual a disponibilidade de equipamentos de proteção individual, de respiradores, de médicos (Política, 2021).

Prosseguindo na análise da "Política de informações enganosas sobre a Covid-19", do *Twitter*, cabe destacar que a plataforma, por sua vez, também esclarece quais são os conteúdos que não representam violação às suas diretrizes. Nesse sentido, ela se posiciona favorável ao "debate público robusto" em relação à pandemia, buscando preservar as manifestações que levem em consideração o conhecimento científico. Desse modo, o *Twitter* determina que serão resguardadas as "opiniões e/ou sátiras fortes, desde que não contenham afirmações falsas ou enganosas sobre o fato". Além disso, a plataforma defende o contra discurso, com a correção de informações inverídicas, bem como discussões sobre os avanços científicos atinentes ao cenário pandêmico, desde que não haja deturpação dos dados obtidos (Política, 2021).

Por conseguinte, o *Twitter* dispõe, ainda, sobre as consequências decorrentes da violação de sua "Política de informações enganosas sobre a Covid-19", mencionando que as medidas a serem adotadas dependem do tipo de transgressão e do histórico de infrações anteriormente cometidas pelo usuário da plataforma. Dentre as ações cabíveis, incluem-se: a exclusão do *tweet*; a marcação do conteúdo violador; e o bloqueio ou a suspensão permanente da conta do usuário (Política, 2021).

Nessa perspectiva, o *Twitter* determina que as violações graves de sua política, como a postagem de informações enganosas relacionadas à origem do Novo Coronavírus e ao tratamento contra a doença, bem como a publicação de teorias conspiratórias sobre a vacina contra a Covid-19, ensejam a remoção do conteúdo, do domínio da plataforma. Destaca-se, entretanto, que a retirada do

tweet ocorre diante do acúmulo de duas transgressões e, além dessa sanção, o usuário fica temporariamente impedido de acessar a sua conta (Política, 2021).

Por outro lado, nas situações que não ensejam a exclusão do conteúdo publicado, o *Twitter* se comprometeu a fornecer "contexto adicional" a *tweets* de teor desinformativo acerca da Covid-19. Nesse sentido, dentre as medidas que podem ser aplicadas, estão: a aplicação de um rótulo à postagem, indicando que esta contém alguma inconsistência; a exibição de um aviso prévio ao compartilhamento ou à curtida de um conteúdo enganoso; a redução da visibilidade de um *tweet*, bem como a desativação de curtidas, respostas e compartilhamentos; e o fornecimento de um link de redirecionamento para a página da "Política de informações enganosas sobre a Covid-19" (Política, 2021).

Por fim, destaca-se que o *Twitter* determina a suspensão de uma conta, caso seja evidenciado que esta se dedica ao compartilhamento de conteúdos equivocados a respeito da Covid-19, estabelecendo, ainda, a proibição da atividade de contas falsas direcionadas à criação de postagens enganosas sobre a pandemia. Além disso, a plataforma estipula a suspensão permanente da conta, em caso de violação grave ou reiterada de suas diretrizes, com base nos seguintes critérios (Política, 2021):

> 1 transgressão: nenhuma ação no nível de conta
> 2 transgressões: bloqueio de conta por 12 horas
> 3 transgressões: bloqueio de conta por 12 horas
> 4 transgressões: bloqueio de conta por 7 dias
> 5 ou mais transgressões: suspensão permanente (Política, 2021).

Diante disso, o *Twitter* esclarece que, em caso de o usuário considerar um equívoco o bloqueio ou a suspensão de sua conta, há a opção de enviar uma contestação à plataforma, para averiguar a situação, a partir de um link existente na própria seção da "Política de informações enganosas sobre a Covid-19" (Política, 2021). Entretanto, em que pese a existência desse "direito ao contraditório", o *Twitter* não fornece explicações a respeito dessa dinâmica, de como é efetuada a análise do caso e de quais são os padrões que determinam ou impedem a reintegração do usuário ao site.

Outro ponto de destaque, nessa perspectiva, é o fato de a "Política de informações enganosas sobre a Covid-19", do *Twitter*, estar em consonância com os estudos e as recomendações das autoridades sanitárias, especialistas da área da saúde e ONGs. Estes parceiros, por sua vez, encontram-se à disposição da plataforma, para prestar informações sobre a conjuntura pandêmica e auxiliar na revisão dos conteúdos potencialmente violadores das regras estabelecidas, por ela própria, visando ao controle da desinformação sobre a Covid-19 em seu domínio (Política, 2021).

Apresentadas as estratégias adotadas pelo *Twitter*, para combater a desinformação sobre a pandemia, entende-se conveniente proceder a uma breve análise geral de sua "Política de informações enganosas sobre a Covid-19". Nesse sentido, primeiramente, sublinha-se que a plataforma agiu no mesmo sentido das demais, estando ciente de que a desinformação virtual é nociva à coletividade,

prejudicando o enfrentamento da crise sanitária global decorrente do Novo Coronavírus, a exemplo da propagação de conteúdos que promovem a ideia da hesitação vacinal, o que compromete a contenção da disseminação da doença.

Assim, reitera-se o argumento de que consiste no papel do próprio *Twitter* atuar na tentativa de combater a desinformação sobre a Covid-19 que circula em seu domínio, por meio do sistema de moderação de conteúdos, pois, conforme mencionado anteriormente, as plataformas não são ambientes neutros e alheios ao que ocorre no mundo *offline* (Gillespie, 2018). Desse modo, acredita-se que a sinalização de conteúdo enganoso, bem como as demais medidas interventivas citadas, tais como a remoção de *tweets* e o bloqueio ou a suspensão de contas, tornam-se importantes para garantir o direito informacional da coletividade, especialmente, no contexto pandêmico (Schreiber, 2017).

Nesse caso, não se pode deixar de comentar que não apenas os internautas são afetados pelo fenômeno da desinformação, mas, também, a população como um todo, haja vista que a circulação dos conteúdos acaba não se restringindo apenas ao *Twitter*, pois estes são difundidos em outros espaços virtuais e não virtuais, podendo atingir a sociedade como um todo, influenciando as suas emoções, as suas ideias e o seu comportamento, colocando em risco a saúde coletiva, no caso da pandemia da Covid-19.

Por outro lado, apesar de a "Política de informações enganosas sobre a Covid-19", do *Twitter*, aparentar constituir uma iniciativa positiva e necessária, destaca-se, entretanto, que existem inúmeros aspectos ocultos, nesse conjunto de diretrizes, que suscitam grandes problemáticas, conforme será comentado a seguir. Nesse sentido, primeiramente, expõe-se que, além da falta de clareza na redação do documento – em decorrência da tradução automática do Inglês para o Português, fornecida pela própria plataforma, o que dificulta a compreensão de determinados pontos – é possível constatar, também, a opacidade de seus termos.

Isso porque a política do *Twitter* se limita a dispor cláusulas genéricas, sem divulgar detalhes acerca da metodologia do sistema de moderação e dos critérios que orientam a aplicação das medidas interventivas às postagens violadoras dos termos da plataforma (exclusão do *tweet*, marcação do conteúdo e bloqueio ou suspensão permanente da conta, conforme exposto anteriormente). Nessa perspectiva, a plataforma acaba não esclarecendo, por exemplo, de que forma os *tweets* são denunciados à moderação (se, exclusivamente, por outros usuários ou se o próprio sistema de inteligência artificial do *Twitter* é que detecta os conteúdos suspeitos de infringir as regras estabelecidas) e por quem é efetuada a análise das postagens denunciadas (se por moderadores humanos ou, exclusivamente, por meio da inteligência artificial), para posterior definição de quais medidas interventivas serão impostas às publicações classificadas como infratoras.

Essa ausência de critérios transparentes, nas cláusulas da política do *Twitter* pode gerar discricionariedades, por parte da plataforma, na ocasião da análise dos *tweets* denunciados, suspeitos de violarem as regras preestabelecidas. Ou seja: existe a possibilidade de conteúdos semelhantes, que tratem sobre um mesmo assunto, não receberem o mesmo tratamento pela plataforma, não lhes sendo aplicadas as mesmas sanções previstas (Rodrigues; Kurtz, 2020). Por conseguinte, evidencia-se que outra polêmica intrínseca à "Política de informações enganosas sobre a Covid-19", do *Twitter*, envolve o questionamento, por parte do público,

quanto à legitimidade da plataforma para aplicar medidas restritivas em relação aos conteúdos, como a exclusão de um *tweet* e o bloqueio ou a suspensão permanente de uma conta. Isso porque tais alternativas poderiam representar o cerceamento do discurso de seus usuários, ocasionando a violação de seu direito à liberdade de expressão (Rodrigues; Kurtz, 2020).

Entretanto, destaca-se que a opção pela não aplicação dessas medidas restritivas – ou seja, a manutenção dos conteúdos desinformativos sobre a Covid-19, no âmbito do *Twitter,* preservando-se, assim, a liberdade de expressão dos internautas – em contrapartida, pode suscitar a violação do direito informacional da coletividade e da proteção à saúde pública, evidenciando-se, então, a possibilidade de instauração de um conflito entre direitos, nesse contexto. Ressalta-se, contudo, que não se pretende ampliar todas essas questões, no presente trabalho, visto que, dada a sua complexidade, merecem uma pesquisa independente, para que sejam explorados todos os pontos os quais a temática possibilita. O fato é que, dentro da abordagem jurídica do tema, possíveis soluções para essas problemáticas podem ser encontradas a partir do Constitucionalismo Digital, que "em seu sentido mais amplo, refere-se à proteção de direitos constitucionais em diversas tecnologias digitais" (Pereira; Keller, 2022, p. 2651).

Por conseguinte, ainda no que tange à "Política de informações enganosas sobre a Covid-19", do *Twitter*, convém expor que a exclusão de determinado *tweet* e o bloqueio ou a suspensão de uma conta – que compreendem as medidas mais radicais, apesar de que, nesses dois últimos casos, o indivíduo pode retornar à ativa, criando outra conta (Gillespie, 2018, p. 176-177) – decorrentes de violações às regras estabelecidas pela plataforma, podem acabar gerando efeitos diversos do esperado. Isso porque, a depender do usuário – especialmente, em se tratando de pessoa pública, de ampla notoriedade – a aplicação de medidas restritivas em relação às suas postagens pode gerar grande repercussão na internet e, até mesmo, no mundo *offline*, atingindo, desse modo, um público maior do que se o conteúdo enganoso e prejudicial tivesse sido mantido apenas no domínio da plataforma.

Exemplo dessa questão é o caso dos *tweets* do Ex-Presidente do Brasil, Jair Bolsonaro, que, em março de 2020, teve uma postagem removida pelo *Twitter*, pois continha um vídeo incitando aglomerações, em meio à pandemia; criticando as medidas de distanciamento social recomendadas pela OMS; e defendendo o uso do fármaco "Cloroquina", para tratamento da Covid-19 (Struck, 2020). Em outra ocasião, em janeiro de 2021, Bolsonaro teve uma postagem marcada, pelo *Twitter*, com um alerta de "informação enganosa e potencialmente prejudicial", por afirmar que o tratamento precoce da Covid-19, por meio de medicamentos antimaláricos, como a "Cloroquina" e a "Hidroxicloroquina" poderiam reduzir a progressão da doença – sendo que não existe comprovação científica acerca da eficácia desses fármacos em relação à Covid-19 (Twitter, 2021).

Destaca-se que todas essas medidas interventivas aplicadas aos *tweets* de Bolsonaro – legitimadas pelos termos da "Política de informações enganosas sobre a Covid-19", do *Twitter* – foram amplamente noticiadas pela imprensa, o que gerou repercussão internacional e ampliou, portanto, o alcance do público ao conteúdo enganoso e prejudicial publicado pelo Presidente. Entretanto, não se pode deixar de comentar que existem divergências, por parte dos pesquisadores, acerca

da efetividade desse tipo de medidas interventivas adotadas pelas plataformas, justamente, diante da possibilidade de elas gerarem o efeito reverso, disseminando os conteúdos desinformativos e prejudicando a coletividade, ao invés de ocultá-los.

Desse modo, pontua-se que os adeptos à adoção dessas estratégias de combate à desinformação – especialmente, a técnica de marcação do conteúdo desinformativo, seguido da correção correspondente, com a demonstração das informações corretas – defendem que se trata do melhor método para eliminar os efeitos da desinformação, reduzindo a crença do público em relação aos conteúdos enganosos, ao mesmo tempo em que há a preservação da liberdade de expressão dos internautas (Avaaz, 2020). Por outro lado, assinala-se que os opositores da aplicação de medidas interventivas, por parte das plataformas, entendem que elas acabam ampliando a propagação dos conteúdos que foram objeto das restrições, gerando prejuízos à coletividade, e, em razão disso, recomenda-se o exercício de uma prática conhecida como "silêncio estratégico"[12] (Tardáguila, 2021, p. 2-4).

Sintetizando todos os desafios intrínsecos ao sistema de moderação do *Twitter*, é possível mencionar o seguinte:

> A partir da atividade de moderação de conteúdo decorrem implicações como limitação da liberdade de expressão, exposição a conteúdos sensíveis, discricionariedade do moderador, controle do discurso público, remoção de conteúdo em massa. Desta feita, perfaz necessário questionar como se afere a legitimidade do conteúdo, de modo a indagar se cabe à plataforma julgar o conteúdo e até que ponto é legítimo a plataforma filtrar o conteúdo e julgar o conteúdo, bem como ter como padrão a remoção do conteúdo (Poletto; Morais, 2022, p. 117).

Por fim, diante de tudo o que foi exposto anteriormente, não se pode negar a importância da "Política de informações enganosas sobre a Covid-19", do *Twitter*, no contexto da crise sanitária global vigente. Entretanto, destaca-se que, para que ela atinja os seus reais objetivos – os quais, em linhas gerais, consistem em combater a desinformação sobre a pandemia, resguardando o direito informacional dos indivíduos e a proteção da saúde coletiva – entende-se necessária a adoção de maior transparência, por parte da plataforma, em relação aos seus usuários, explicitando, principalmente, a metodologia da moderação de conteúdos, bem como os critérios que orientam a aplicação das intervenções às postagens violadoras de suas diretrizes.

---

[12] De forma breve, destaca-se que o "silêncio estratégico" consiste em uma medida bastante tradicional, pela qual a mídia omite assuntos potencialmente danosos, a fim de evitar que mensagens prejudiciais se espalhem pela comunidade. Essa estratégia pode ser adotada, também, no contexto das plataformas de conteúdo gerado pelos internautas, tanto por parte dos meios de comunicação, quanto por parte do público em geral, especialmente, a fim de impedir que conteúdos desinformativos e violadores dos direitos de terceiros sejam propagados no âmbito virtual, gerando efeitos prejudiciais aos indivíduos e à coletividade. Nessa perspectiva, como forma de exercer essa estratégia, é recomendável que os internautas não curtam, comentem, compartilhem e deem engajamento à desinformação e a manifestações ofensivas e, do mesmo modo, que os veículos de comunicação evitem reverberar esses conteúdos ou pautas relacionadas a eventuais medidas restritivas que lhes sejam impostas, pelas plataformas (InternetLab, 2021).

De acordo com Gillispie (2018, p. 199), a transparência das plataformas não supõe apenas a "ausência de opacidade", mas sim tornar os procedimentos atrelados à moderação visíveis aos usuários, ainda que de forma discreta. Esclarece-se que a transparência na moderação de conteúdo do *Twitter*, na conjuntura pandêmica, é essencial para a preservação e a harmonização de direitos e garantias fundamentais – que podem acabar sendo violados, com a concretização das medidas previstas pela sua "Política de informações enganosas sobre a Covid-19", conforme demonstrado na presente seção.

CONCLUSÃO

A partir de tudo o que se expôs no presente trabalho, foi possível constatar que a desinformação consiste em um dos grandes problemas que estão sendo enfrentados na atual conjuntura, gerando graves consequências a todos os âmbitos da sociedade em rede – como a saúde pública e a democracia. Evidenciou-se que o fenômeno em questão é provocado por causas econômicas, políticas e ideológicas, tendo a sua propagação facilitada no ciberespaço, especialmente, pela arquitetura das plataformas de conteúdo gerado pelos usuários.

Destacou-se, ainda, que a desinformação virtual foi agravada no contexto da crise sanitária global provocada pela Covid-19, violando, então, o direito informacional dos indivíduos e expondo a saúde coletiva a risco, tendo em vista a circulação desenfreada de informações enganosas e prejudiciais, envolvendo a pandemia, nas plataformas digitais de conteúdo gerado pelo usuário. Atentas a essa perspectiva e cientes de seu papel enquanto vetores de conteúdos desinformativos, expôs-se que as plataformas adotaram determinadas medidas para combater a "infodemia" e, consequentemente, contribuir para a redução de danos no mundo *offline*.

Diante disso, buscou-se verificar quais as estratégias adotadas pelo *Twitter*, para combater a desinformação sobre a Covid-19 que é propagada em seu domínio, no contexto da pandemia, bem como os desafios e perspectivas relacionados ao seu sistema de moderação. Assim, por meio da análise de sua "Política de informações enganosas sobre a Covid-19", constatou-se que a plataforma em questão proíbe, como regra geral, a circulação de conteúdos falsos sobre a conjuntura pandêmica que possam causar danos aos indivíduos. Além disso, apurou-se que, como forma de sanção ao descumprimento das diretrizes, o *Twitter* estabelece determinadas categorias de medidas interventivas a serem aplicadas às postagens "transgressoras", a saber: a exclusão do *tweet*; a marcação do conteúdo violador; e o bloqueio ou a suspensão permanente da conta do usuário.

A partir dessa investigação, evidenciou-se a impossibilidade de se averiguar a efetividade prática da "Política de informações enganosas sobre a Covid-19", do *Twitter*, eis que isso demandaria um estudo mais complexo. O que se pretendeu, no presente trabalho, contudo, foi demonstrar que a empresa está atenta aos problemas sociais, buscando atender às demandas que estão relacionadas à sua área de atuação, visto que as plataformas não são espaços neutros e desprovidos de regulamentação interna, exercendo, portanto, a sua governança por meio da autorregulação. Assim, não restam dúvidas quanto à legitimidade tida, pelo *Twitter*, em aplicar as medidas interventivas decorrentes do sistema de moderação de conteúdos, sobretudo, a suspensão de conta, pois a plataforma constitui um

negócio privado, regido pelas próprias regras, sendo que os usuários consentem com elas, no momento de sua adesão ao *site*[13].

Entretanto, também foi possível constatar que, apesar de aparentar constituir uma iniciativa exemplar, observou-se que a referida política do *Twitter* compreende inúmeros pontos controversos, especialmente, no que tange à falta de transparência de suas cláusulas – o que pode ensejar a violação e a colisão de direitos, como é o caso da liberdade de expressão, do direito informacional e da proteção à saúde pública. Nessa perspectiva, destacou-se, ainda, que o estabelecimento de diretrizes mais claras, devidamente delineadas e de ampla publicidade, aos usuários da plataforma, constitui uma alternativa para garantir os direitos fundamentais dos internautas e da coletividade, no contexto da "infodemia" do Novo Coronavírus.

Tudo isso é reflexo da falta de transparência do próprio sistema de moderação de conteúdos do *Twitter*, que deveria esclarecer aos usuários sobre os procedimentos que orientam a aplicação de medidas interventivas em relação às postagens violadoras das diretrizes da plataforma, de forma a manter uma relação de equilíbrio entre a empresa e o internauta, eis que a relação estabelecida entre ambos pode ser considerada de consumo, sendo, portanto, aquele último, o mais vulnerável, diante do poderio da *big tech*.

Além das medidas elencadas anteriormente, voltadas ao enfrentamento da desinformação pandêmica, entende-se que a alternativa mais coerente que deveria ser adotada pelo *Twitter* é o aperfeiçoamento de seu sistema de moderação de conteúdos, especialmente, no que tange à aplicação da medida interventiva de marcação do conteúdo violador de suas diretrizes. Assim, ao detectar um post desinformativo a respeito da Covid-19, a plataforma, além de demonstrar que o conteúdo em questão é irregular, poderia fornecer a correção das informações constantes no *tweet*, impulsionando-as, a fim de que sejam de amplo conhecimento dos usuários da plataforma. Aparentemente, trata-se de uma opção viável, sobretudo, do ponto de vista da tentativa de harmonização entre os direitos fundamentais dos internautas e da coletividade, como é o caso da liberdade de expressão do usuário do *Twitter* e a proteção da saúde pública e o direito informacional da coletividade, em tempos pandêmicos.

Por fim, não se pode deixar de comentar que todas essas questões relacionadas à autorregulação das plataformas em geral e, de forma específica, à moderação de conteúdos do *Twitter*, são complexas e não pressupõem uma solução única e simplista para os seus problemas. O ideal é que a plataforma esteja sempre atenta às realidades e às necessidades da sociedade como um todo – o que reflete, diretamente, em seu meio de atuação, a exemplo da desinformação sobre a pandemia da Covid-19 – e que esteja sempre em diálogo com a comunidade científica e com os especialistas da área, para aperfeiçoar os seus sistemas e

[13] Há que se deixar claro, no entanto, que, nesta circunstância, a legitimidade da plataforma não a isenta de cometer equívocos na atividade moderadora, havendo a remoção injusta, por exemplo, de uma postagem que não viola os termos de uso do site. Logo, apesar dessa legitimidade, que não é "absoluta", nos casos em que forem identificados supostos erros na moderação, o Twitter aceita o recebimento de contestação, por parte do usuário prejudicado, visando à revisão e reintegração do conteúdo excluído (Nossas, 2023).

diretrizes internas. Tudo isso, a fim de proporcionar uma experiência satisfatória e equilibrada para seus usuários, anunciantes e futuros membros.

REFERÊNCIAS

AGRELA, L. Como as redes sociais estão combatendo a desinformação sobre o coronavírus. *Exame*. [S.L.], p. 1-8. 30 mar. 2020. Disponível em: https://exame.com/tecnologia/como-as-redes-sociais-estao-combatendo-a-desinformacao-sobre-o-coronavirus/. Acesso em: 20 jul. 2021.

ALMEIDA, C. L. de. *Regulação da transparência em plataformas digitais e legitimidade na moderação de conteúdo*. 2022. 134 f. Dissertação (Mestrado) - Curso de Mestrado em Direito, Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas, Rio de Janeiro, 2022. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/32486. Acesso em: 23 fev. 2023.

ALVES, M. A. S.; MACIEL, E. R. H. O fenômeno das fake news: definição, combate e contexto. *Internet&Sociedade*, [S. L.], 2020, n.1, v. 1, p. 144-171, fev. 2020. Disponível em: https://revista.internetlab.org.br/o-fenomeno-das-fake-news-definicao-combate-e-contexto/. Acesso em: 25 maio. 2021.

AVAAZ. *Princípios Legislativos para Combater a Desinformação*. [S.L.], p. 1-8. Maio 2019. Disponível em: https://secure.avaaz.org/campaign/po/disinfo_legislative_principles/. Acesso em: 25 maio. 2021.

BRASIL. Lei n. 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. *Diário Oficial da União*. Poder Legislativo, Brasília, DF, 24 abr. 2014. p. 1. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 25 jul. 2021.

BRITES, M. J.; AMARAL, I.; CATARINO, F. A era das "fake news": o digital storytelling como promotor do pensamento crítico. *Journal of Digital Media & Interaction*, [S.L.], 2018, n. 1, v. 1, p. 85-98, 2018, Disponível em: https://hdl.handle.net/1822/55530. Acesso em: 01 jul. 2021.

CASTELLS, M. *A galáxia da internet*: reflexões sobre a internet, os negócios e a sociedade. Rio de Janeiro: Jorge Zahar, 2003.

CASTELLS, M. *O poder da comunicação*. 1. ed. São Paulo/Rio de Janeiro: Paz e Terra, 2015.

COMM, J. *O poder do Twitter*: estratégias para dominar seu mercado e atingir seus objetivos com um tweet por vez. São Paulo: Editora Genta, 2009.

COUTO, N. de M. *O papel regulatório do Estado na moderação de conteúdo exercida pelas plataformas de redes sociais*. 2022. 234 f. Dissertação (Mestrado) - Curso de Mestrado em Direito, Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas, Rio de Janeiro, 2022. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/33008. Acesso em: 22 fev. 2023.

DEODATO, P. G. O.; SOUSA, A. Fake news e o processo de impeachment de Dilma Rousseff: uma análise de notícias falsas publicadas pelo site "Pensa Brasil". *Temática*, [S. L.], 2018, n. 11, v. 14, p. 109-124, nov. 2018. Disponível em: https://doi.org/10.22478/ufpb.1807-8931.2018v14n11.42954. Acesso em: 01 jul. 2021.

GILLESPIE, T. *Custodians of the internet*: platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018.

GORWA, R. What is platform governance? *Information, Communication & Society*, Oxford, v. 22, n. 6, p. 1-29, 11 fev. 2019. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1573914?journalCode=rics20. Acesso em: 24 fev. 2023.

GRIMMELMANN, J. The virtues of moderation. Yale JL & Tech., v. 17, 2015. Disponível em: https://scholarship.law.cornell.edu/facpub/1486/#:~:text=This%20Article%20provides%20a%20novel,sterility%20of%20too%20much%20control. Acesso em: 23 fev. 2023.

INTERNETLAB. *Falando sobre ataques online e trolls*: um guia para jornalistas e criadores de conteúdo na internet. [S.L.], p. 1-30. 2021. Disponível em: https://www.internetlab.org.br/wp-content/uploads/2021/06/guia_trolls_paginasimples_12062021_ok.pdf. Acesso em: 10 ago. 2021.

KURTZ, L.; DO CARMO, P. R. R.; VIEIRA, V. B. R. *Transparência na moderação de conteúdo*: tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: https://bit.ly/3xjAUka. Acesso em: 06 jul. 2021.

LÉVY, P. *Ciberdemocracia*. Lisboa: Instituto Piaget, 2002.

MANJOO, F. *True Enough*: Learning to live in a post-fact society. John Wiley & Sons: New Jersey, 2008.

NOSSAS opções de medidas corretivas. *Twitter*. 2023. Disponível em: https://help.twitter.com/pt/rules-and-policies/enforcement-options#:~:text=Mesmo%20que%20uma%20conta%20esteja,dependendo%20da%20natureza%20da%20viola%C3%A7%C3%A3o. Acesso em: 13 mar. 2023.

O BRASIL está sofrendo uma infodemia de Covid-19. *Avaaz*. [S.L.], p. 1-14. 04 maio 2020. Disponível em: https://secure.avaaz.org/campaign/po/brasil_infodemia_coronavirus/. Acesso em: 25 maio. 2021.

OMS considera coronavírus 'maior crise sanitária mundial da nossa época'. *Uol*. São Paulo. 16 mar. 2020. Disponível em: https://noticias.uol.com.br/ultimasnoticias/afp/2020/03/16/oms-considera-coronavirus-maior-crise-sanitaria-mundial-da-nossaepoca.htm. Acesso em: 25 jul. 2021.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. *Entenda a infodemia e a desinformação na luta contra a COVID-19*. [S.L.], p. 1-5. 2020. Disponível em: https://iris.paho.org/bitstream/handle/10665.2/52054/Factsheet-Infodemic_por.pdf?sequence=14. Acesso em: 25 maio. 2021.

PARISER, E. *O filtro invisível*: o que a internet está escondendo de você. Rio de Janeiro: Zahar, 2012.

PEREIRA, J. R. G.; KELLER, C. I. Constitucionalismo Digital: contradições de um conceito impreciso. *Revista Direito e Práxis*, Rio de Janeiro, v. 13, n. 4, p. 2648-2689, 23 out. 2022. Disponível em: https://www.scielo.br/j/rdp/a/5bpy8smKHgXbKqKzDWDCZQm/#. Acesso em: 25 fev. 2023.

POLETTO, Á. E.; MORAIS, F. S. de. A moderação de conteúdo em massa por plataformas privadas de redes sociais. *Prisma Jurídico*, São Paulo, v. 21, n. 1, p. 108-126, jan./jun. 2022. http://doi.org/10.5585/prismaj.v21n1.20573. Disponível em: https://periodicos.uninove.br/prisma/article/view/20573/9644. Acesso em: 17 fev. 2023.

POLÍTICA de informações enganosas sobre a COVID-19. *Twitter*. [S.L.] 2021. Disponível em: https://help.twitter.com/pt/rules-and-policies/medical-misinformation-policy. Acesso em: 25 jul. 2021.

RIBEIRO; M. M.; ORTELLADO, P. O que são e como lidar com as notícias falsas. *SUR - Revista Internacional de Direitos Humanos*, [S. L.], 2018, n.27, v. 15, p. 71-83, 2018. Disponível em: sur-27-portugues-marcio-moretto-ribeiro-pablo-ortellado.pdf (conectas.org). Acesso em: 01 jul. 2021.

RODRIGUES, G.; KURTZ, L. *Transparência sobre moderação de conteúdo em políticas de comunidade*. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2020. Disponível em: https://irisbh.com.br/publicacoes/transparencia-sobre-moderacao-de-conteudo-em-politicas-de-comunidade/. Acesso em: 25 maio. 2021.

SCHREIBER, A. Art. 5º, IX. *In*: MORAES, A. de, *et al.*. *Constituição Federal Comentada*. Rio de Janeiro: Forense, 2018. p. 64- 66.

SHANE, T. The psychology of misinformation: Why we're vulnerable. *First Draft*. 30 jun. 2020. Disponível em: https://firstdraftnews.org/articles/the-psychology-of-misinformation-why-were-vulnerable/. Acesso em: 08 mar. 2023.

SPINELLI, E. M.; SANTOS, J. de A. Jornalismo na era da pós-verdade: Fact-checking como ferramenta de combate às fake news. *Revista Observatório*, [S. L.], 2018, v. 4, n. 3, p. 759-782, 29 abr. 2018. Disponível em: https://sistemas.uft.edu.br/periodicos/index.php/observatorio/article/view/4629. Acesso em: 01 jul. 2021.

STRUCK, J. Com tuítes apagados, Bolsonaro se junta a Maduro e aiatolá do Irã. *Uol*. [S.L]. 23 mar. 2020. Disponível em: https://noticias.uol.com.br/ultimas-noticias/deutschewelle/2020/03/30/com-tuites-apagados-bolsonaro-se-junta-a-maduro-e-aiatola-do-ira.htm?cmpid=copiaecola. Acesso em: 25 jul. 2021.

TARDÁGUILA, C. #SilêncioEstratégico: um forte antídoto para Venturinis e Sikêras Jr. *Uol*. [S.L]. 2021. Disponível em: https://noticias.uol.com.br/colunas/cristina-tardaguila/2021/06/29/venturini-sikera-covid-homofobia-silencio-estrategico.htm. Acesso em: 20 jul. 2021.

TAYLOR, S. *The Psychology of Pandemics*: Preparing for the Next Global Outbreak of Infectious Disease. Newcastle: Cambridge Scholars Publishing, 2019.

TEFFÉ, C. S. de. Fake news: como proteger a liberdade de expressão e inibir notícias falsas? *ITS Rio*. [S.L]. 19 mar. 2018. Disponível em: https://feed.itsrio.org/fake-newscomo-proteger-a-liberdade-deexpress%C3%A3o-e-inibir-not%C3%ADcias-falsas8058aedd9f5c. Acesso em: 25 maio. 2021.

TIC DOMICÍLIOS. *Pesquisa Sobre o Uso das Tecnologias de Informação e Comunicação nos Domicílios Brasileiros*. São Paulo: Comitê Gestor da Internet no Brasil, 2019. Disponível em: https://www.cetic.br/pesquisa/domicilios/publicacoes/. Acesso em: 25 maio. 2021.

TWITTER põe alerta em post de Bolsonaro sobre tratamento precoce da covid. *Uol*. [S.L]. 15 jan. 2021. Disponível em: https://www.uol.com.br/tilt/noticias/redacao/2021/01/15/twitter-poe-alerta-em-post-de-bolsonaro-sobre-tratamento-precoce-da-covid.htm. Acesso em: 25 maio. 2021.

VOLPATO, B. Ranking: as redes sociais mais usadas no Brasil e no mundo em 2022, com insights e materiais. *Resultados Digitais*. 23 maio. 2022. Disponível em: https://resultadosdigitais.com.br/marketing/redes-sociais-mais-usadas-no-brasil/. Acesso em: 06 mar. 2023.

WARDLE, C. First Draft's essential guide to understanding information disorder. *First Draft*. [S.L]. Out. 2019a. Disponível em: https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x76701. Acesso em: 25 jul. 2020.

WARDLE, C. Information disorder: 'The techniques we saw in 2016 have evolved'. *First Draft*. [*S.l.*]. 21 out. 2019b. Disponível em: https://firstdraftnews.org/latest/information-disorder-thetechniques-we-saw-in2016-have-evolved/. Acesso em: 25 maio. 2021.

WU, T. *The Attention Merchants*: The Epic Scramble to Get Inside Our Heads. New York: Knopf, 2016.

ZUBOFF, S. *A Era do Capitalismo de Vigilância*: a luta por um futuro humano na nova fronteira de poder. Rio de Janeiro: Intrínseca, 2021